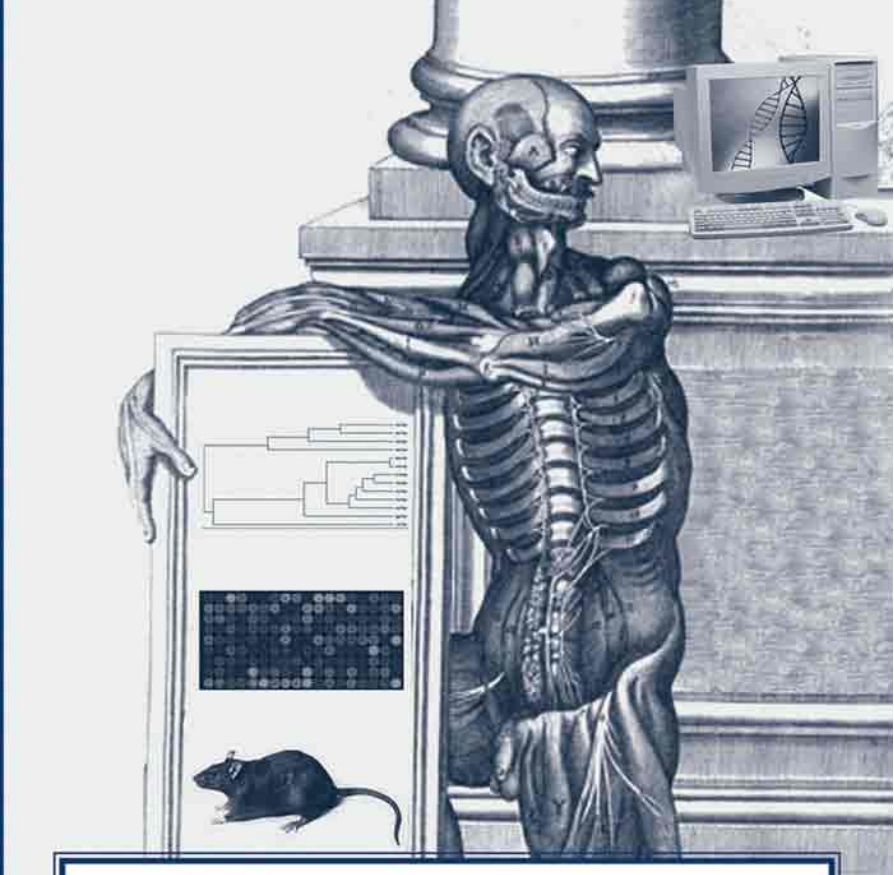


BIOINFORMATICS and FUNCTIONAL GENOMICS

JONATHAN PEVSNER

BIOINFORMATICS AND FUNCTIONAL GENOMICS

This Page Intentionally Left Blank

Bioinformatics and Functional Genomics

This Page Intentionally Left Blank

BIOINFORMATICS AND FUNCTIONAL GENOMICS

Jonathan Pevsner

Department of Neurology,
Kennedy Krieger Institute

and

Department of Neuroscience and
Division of Health Sciences Informatics,
The Johns Hopkins School of Medicine
Baltimore, Maryland

WILEY-LISS
A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

The cover image is modified from plate X of *Tabulae anatomicae* by Pietro Berrettini (Rome, 1741). The original image was provided by the Institute of the History of Medicine, The Johns Hopkins University. Used with permission.

Library of Congress Cataloging-in-Publication Data:
Pevsner, Jonathan, 1961–
Bioinformatics and functional genomics / Jonathan
Pevsner.
p. ; cm.
Includes bibliographical references and index.
ISBN 0-471-21004-8 (paper : alk. paper)
1. Genomics. 2. Bioinformatics. 3. Proteomics. 4. Genetics—Data processing.
[DNLM: 1. Computational Biology—methods. 2. Genomics. 3. Genetic
Techniques. 4. Genome, Human. 5. Proteome. QU 26.5 P514b 2003] I. Title.
QH441.2.P48 2003
572.8'6—dc21
2002156139

Printed in the United States of America

10 9 8 7 6 5 4 3 2

To Madeline Cheshire

This Page Intentionally Left Blank

Contents in Brief

PART I ANALYZING DNA, RNA, AND PROTEIN SEQUENCES IN DATABASES

1 Introduction, 3

2 Access to Sequence Data and Literature Information, 15

3 Pairwise Sequence Alignment, 41

4 Basic Local Alignment Search Tool (BLAST), 87

5 Advanced BLAST Searching, 127

PART II GENOMEWIDE ANALYSIS OF RNA AND PROTEIN

6 Bioinformatic Approaches to Gene Expression, 157

7 Gene Expression: Microarray Data Analysis, 189

8 Protein Analysis and Proteomics, 223

9 Protein Structure, 273

10 Multiple Sequence Alignment, 319

11 Molecular Phylogeny and Evolution, 357

PART III GENOME ANALYSIS

12 Completed Genomes and the Tree of Life, 397

13 Completed Genomes: Viruses, 437

14 Completed Genomes: Bacteria and Archaea, 465

15 Eukaryotic Genomes: Fungi, 503

16 Eukaryotic Genomes: From Parasites to Primates, 539

17 Human Genome, 607

18 Human Disease, 647

This Page Intentionally Left Blank

Contents

Foreword, xix

Preface, xxiii

PART I ANALYZING DNA, RNA, AND PROTEIN SEQUENCES IN DATABASES

1 **Introduction, 3**
Organization of The Book, 4
Bioinformatics: The Big Picture, 4
A Consistent Example: Retinol-Binding Protein, 7
Organization of The Chapters, 9
Suggestions For Students and Teachers: Web Exercises and Find-a-Gene, 9
Key Bioinformatics Websites, 10
Suggested Reading, 12
References, 12

2 **Access to Sequence Data and Literature Information, 15**
Introduction to Biological Databases, 15
GenBank: Database of Most Known Nucleotide and Protein Sequences, 16
Amount of Sequence Data, 17
Organisms in GenBank, 17
Types of Data in GenBank, 19
Genomic DNA Databases, 19
cDNA Databases Corresponding to Expressed Genes, 20
Protein Databases, 20
Expressed Sequence Tags (ESTs), 20
ESTs and UniGene, 21
Sequence-Tagged Sites (STSs), 22
Genome Survey Sequences (GSSs), 22
High-Throughput Genomic Sequence (HTGS), 23
National Center for Biotechnology Information, 24
Introduction to NCBI: Home Page, 24
PubMed, 24
Entrez, 24
BLAST, 25
OMIM, 25
Books, 26
Taxonomy, 26
Structure, 26
Accession Numbers: Labels to Identify Sequences, 26
Box 2-1. Types of Accession Numbers, 27
Five Ways to Access DNA and Protein Sequences, 27
(1) LocusLink: Centralized Resource for Information on Genes and Proteins, 27
NCBI Reference Sequence (RefSeq): Best Representative Sequences, 29
(2) UniGene, 31
(3) Entrez, 31
(4) European Bioinformatics Institute and Ensembl, 31
(5) ExPASy, 33
Example of How to Access Sequence Data: HIV *pol*, 33

Access to Biomedical Literature, 35
PubMed Central and Movement toward Free Journal Access, 35
Example of PubMed Search: RBP, 35
Box 2-2. Venn Diagrams of Boolean Operators AND, OR, and NOT for Hypothetical Search Terms 1 and 2, 36
Perspective, 36
Pitfalls, 37
Web Resources, 37
Discussion Questions, 37
Problems, 37
Self-Test Quiz, 38
Suggested Reading, 39
References, 39

3 **Pairwise Sequence Alignment, 41**
Introduction, 41
Protein Alignment: Often More Informative Than DNA Alignment, 41
Definitions: Homology, Similarity, Identity, 42
Box 3-1. Structures and One- and Three-Letter Abbreviations of Twenty Common Amino Acids, 45
Box 3-2. Algorithms and Programs, 46
Gaps, 47
Pairwise Alignment, Homology, and Evolution of Life, 47
Dayhoff Model: Accepted Point Mutations, 50
Box 3-3. Dayhoff's Protein Superfamilies, 50
PAM1 Matrix, 51
PAM250 and Other PAM Matrices, 53
Box 3-4. Genetic Code, 54
Box 3-5. Matrix Multiplication, 56
From a Mutation Probability Matrix to a Log-Odds Score Matrix, 57
Practical Usefulness of PAM Matrices in Pairwise Alignment, 59
Important Alternative to PAM: BLOSUM Scoring Matrices, 60
Pairwise Alignment and Limits of Detection, 61
Alignment Algorithms: Global and Local, 62
Global Sequence Alignment: Algorithm of Needleman and Wunsch, 63
Box 3-6. Genetics Computer Group, 68
Local Sequence Alignment: Smith and Waterman Algorithm, 69
Rapid, Heuristic Versions of Smith–Waterman: FASTA and BLAST, 71
Basic Local Alignment Search Tool (BLAST), 72
Significance of Pairwise Alignments: Percent Identity, 73
Box 3-7. Dot Plots, 75
Tests for Statistical Significance of Pairwise Alignments, 75
Statistical Significance of Global Alignments, 76
Statistical Significance of Local Alignments, 77
Perspective, 77
Pitfalls, 78
Web Resources, 78
Discussion Questions, 79
Problems, 80
Self-Test Quiz, 82
Suggested Reading, 83
References, 84

4 **Basic Local Alignment Search Tool (BLAST), 87**
Introduction, 87
BLAST Search Steps, 89
Step 1: Specifying Sequence of Interest, 89
Step 2: Selecting BLAST Program, 90
Step 3: Selecting a Database, 92
Step 4a: Selecting Optional Search Parameters, 92

Step 4b: Selecting Optional Formatting Parameters, 97
BLAST Algorithm Uses Local Alignment Search Strategy, 100
BLAST Algorithm Parts: List, Scan, Extend, 101
BLAST Algorithm: Local Alignment Search Statistics and *E* Value, 103
Making Sense of Raw Scores with Bit Scores, 106
BLAST Algorithm: Relation between *E* and *P* Values, 106
Gapped Alignments in BLAST, 107
Getting to Bottom of BLAST Search, 107
BLAST Search Strategies, 108
General Concepts, 108
Principles of BLAST Searching, 108
How to Evaluate Significance of Your Results, 108
How to Handle too Many Results, 113
How to Handle too Few Results, 113
BLAST Searching with Multidomain Protein: HIV-1 pol, 114
BLAST Searching with Lipocalins: Effect of Changing Scoring Matrices, 117
Perspective, 119
Pitfalls, 119
Web Resources, 122
Discussion Questions, 122
Problems, 123
Self-Test Quiz, 124
Suggested Reading, 124
References, 125

5 **Advanced BLAST Searching, 127**
Introduction, 127
Specialized BLAST Sites, 127
Organism-Specific BLAST Sites, 128
Ensembl BLAST, 130
The Institute for Genomic Research (TIGR) BLAST, 130
Molecule-Specific BLAST Sites, 133
Specialized BLAST Servers and BLAST-Related Algorithms, 133
BLAST-Like Alignment Tools to Search Genomic DNA Rapidly, 135
Finding Distantly Related Proteins: Position-Specific Iterated BLAST (PSI-BLAST), 137
Assessing Performance of PSI-BLAST, 143
PSI-BLAST Errors: Problem of Corruption, 144
Pattern-Hit Initiated BLAST (PHI-BLAST), 145
Using BLAST for Gene Discovery, 147
Perspective, 150
Pitfalls, 151
Web Resources, 151
Discussion Questions, 152
Problems, 153
Self-Test Quiz, 153
Suggested Reading, 154
References, 154

PART II GENOMEWIDE ANALYSIS OF RNA AND PROTEIN

6 **Bioinformatic Approaches to Gene Expression, 157**
Introduction, 157
mRNA: Subject of Gene Expression Studies, 160
Analysis of Gene Expression in cDNA Libraries, 162
Pitfalls in Interpreting Expression Data from cDNA Libraries, 166
TIGR Gene Indices, 169
Serial Analysis of Gene Expression (SAGE), 169
Microarrays: Genomewide Measurement of Gene Expression, 172
Stage 1: Experimental Design for Microarrays, 177
Stage 2: RNA Preparation and Probe Preparation, 178

Stage 3: Hybridization of Labeled Samples to DNA Microarrays, 178
Stage 4: Image Analysis, 179
Stage 5: Data Analysis, 180
Stage 6: Biological Confirmation, 182
Stage 7: Microarray Databases, 182
Stage 8: Further Analyses, 182
Perspective, 183
Pitfalls, 183
Web Resources, 184
Discussion Questions, 185
Problems, 185
Self-Test Quiz, 185
Suggested Reading, 186
References, 186

7 **Gene Expression: Microarray Data Analysis, 189**
Introduction, 189
Microarray Data Analysis: Preprocessing, 191
Global Normalization, 192
Scatter Plots, 193
Box 7-1. Analyzing Data from a Microarray Experiment, 194
Local versus Global Normalization, 194
Box 7-2. Data Transformations to Make Enhanced Scatter Plots, 198
Microarray Data Analysis: Inferential Statistics, 198
Significance Analysis of Microarrays (SAM), 200
Microarray Data Analysis: Descriptive Statistics, 203
Box 7-3. Euclidean Distance, 203
Box 7-4. Pearson Correlation Coefficient *r*, 204
Hierarchical Cluster Analysis of Microarray Data, 204
Partitioning Methods for Clustering: *k*-Means Clustering, 210
Clustering Strategies: Self-Organizing Maps, 210
Principal Components Analysis: Visualizing Microarray Data, 211
Supervised Data Analysis for Classification of Genes or Samples, 213
Annotation of Microarray Data, 214
Perspective, 214
Pitfalls, 215
Web Resources, 216
Discussion Questions, 217
Problems, 217
Self-Test Quiz, 219
Suggested Reading, 220
References, 220

8 **Protein Analysis and Proteomics, 223**
Introduction, 223
Approaches to Proteins: Four Perspectives, 223
Perspective 1. Protein Domains and Motifs: Modular Nature of Proteins, 225
Added Complexity of Multidomain Proteins, 229
Protein Patterns: Motifs or Fingerprints Characteristic of Proteins, 231
Perspective 2. Physical Properties of Proteins, 233
Introduction to Perspectives 3 and 4: Gene Ontology Consortium, 237
Perspective 3: Protein Localization, 242
Perspective 4: Protein Function, 243
Proteomics: Bioinformatic Tools for High-Throughput Protein Analysis, 247
Two-Dimensional Gels Electrophoresis, 248
Box 8-1. Protein Sequencing by Edman Degradation, 250
Box 8-2. Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Spectroscopy MALDI-TOF, 251

Affinity Chromatography and Mass Spectrometry, 252
Yeast Two-Hybrid System, 254
Box 8-3. Yeast Two-Hybrid System, 255
Rosetta Stone Approach, 258
Bioinformatic Approaches to Cellular Pathways, 258
Perspective, 263
Pitfalls, 264
Web Resources, 265
Discussion Questions, 267
Problems, 268
Self-Test Quiz, 268
Suggested Reading, 269
References, 269

9 **Protein Structure, 273**
Overview of Protein Structure and Structural Genomics, 273
Protein Structure, Homology, and Functional Genomics, 274
Biological Questions Addressed by Structural Genomics: Lipocalins, 276
Principles of Protein Structure: From Primary to Secondary Structure, 276
Box 9-1. X-Ray Crystallography, 278
Tertiary Protein Structure: Protein-Folding Problem, 281
Experimental Approaches to Protein Structure, 283
Target Selection and Acquisition of Three-Dimensional Protein Structures, 285
Protein Data Bank, 287
Accessing PDB Entries at NCBI Website, 289
Integrated Views of Universe of Protein Folds, 292
Taxonomic System for Protein Structures: SCOP Database, 293
Box 9-2. VAST Information, 294
CATH Database, 297
Dali Domain Dictionary, 301
FSSP Database, 302
Computational Biology Approaches to Structure, 303
Comparative Modeling, 303
Ab Initio Prediction, 309
Protein Structure Prediction and Limits of Protein Fold Space, 310
Protein Structure and Disease, 311
Perspective, 312
Pitfalls, 312
Web Resources, 313
Discussion Questions, 313
Problems, 313
Self-Test Quiz, 314
Suggested Reading, 315
References, 315

10 **Multiple Sequence Alignment, 319**
Introduction, 319
Definition of Multiple Sequence Alignment, 320
Typical Uses and Practical Strategies of Multiple Sequence Alignment, 321
Feng and Doolittle's Progressive Sequence Alignment, 321
From Multiple Sequence Alignment to Profile Hidden Markov Models, 325
Box 10-1. Similarity Versus Distance Measures, 326
Two Multiple Sequence Alignment Programs, 329
Databases of Multiple Sequence Alignments, 331
Pfam: Protein Family Database of Profile HMMs, 331
SMART, 332
Conserved Domain Database, 333
BLOCKS, 334
PRINTS, 334
PROSITE, 335
Integrated Multiple Sequence Alignment Resources: InterPro, MetaFam, iProClass, 336
PopSet, 340

Multiple Sequence Alignment Database Curation: Manual Versus Automated, 341
User-Generated Multiple Sequence Alignments, 341
ClustalW and ClustalX, 343
PileUp (GCG), 345
Other Multiple Sequence Alignment Software, 345
Assessment of Alternative Multiple Sequence Alignment Algorithms, 346
Perspective, 350
Pitfalls, 351
Web Resources, 351
Discussion Questions, 351
Problems, 351
Self-Test Quiz, 352
Suggested Reading, 353
References, 353

11 **Molecular Phylogeny and Evolution, 357**
Introduction to Molecular Evolution, 357
Historical Background, 358
Molecular Clock Hypothesis, 360
Neutral Theory of Molecular Evolution, 363
Box 11-1. Rate of Nucleotide Substitution *r* and Time of Divergence *T*, 364
Goals of Molecular Phylogeny, 364
Molecular Phylogeny: Nomenclature of Trees, 365
Tree Roots, 368
Enumerating Trees, 368
Box 11-2. Number of Rooted and Unrooted Trees, 369
Species Trees Versus Gene/Protein Trees, 369
Four Stages of Phylogenetic Analysis, 370
Stage 1: Molecular Phylogeny Can Be Performed with DNA, RNA, or Protein, 371
Stage 2: Multiple Sequence Alignment, 375
Stage 3: Tree-Building Methods, 377
Making Trees Using Distance-Based Methods, 378
Making Trees Using UPGMA Distance-Based Method, 379
Box 11-3. Trees from Microarray Data Versus Phylogenetic Trees, 383
Making Trees by Distance-Based Methods: Neighbor Joining (NJ), 383
Making Trees by Character-Based Methods, 383
Maximum Parsimony (MP), 383
Maximum Likelihood (ML), 386
Stage 4: Evaluating Trees Using Randomizing Tests and Bootstrapping, 386
Perspective, 389
Pitfalls, 389
Web Resources, 389
Discussion Question, 389
Problem, 389
Self-Test Quiz, 391
Suggested Reading, 392
References, 392

PART III GENOME ANALYSIS

12 **Completed Genomes and the Tree of Life, 397**
Introduction, 397
Brief History of Systematics, 399
History of Life on Earth, 399
Molecular Sequences as Basis of Tree of Life, 401
Role of Bioinformatics in Taxonomy, 402
Genome-Sequencing Projects: Historical Overview, 404
Brief Chronology of Genome-Sequencing Projects, 404
Overview: 1977–Present, 404
First Viral Genome (1977), 406
First Eukaryotic Organellar Genome (1981), 407

First Chloroplast Genomes (1986), 408
First Eukaryotic Chromosome (1992), 408
Complete Genome of Free-Living Organism (1995), 409
First Eukaryotic Genome (1996), 413
More Bacteria and Archaea (1997), 413
First Genome of Multicellular Organism (1998), 414
Human Chromosome (1999), 414
Fly, Plant, and Human Chromosome 21 (2000), 415
Draft Sequences of Human Genome (2001), 415
Continuing Rise in Completed Genomes (2002), 417
Overview of Genome Analysis, 418
Selection of Genome for Sequencing, 418
Should an Individual from a Species, Several Individuals, or Many Individuals be Sequenced?, 419
How Big Are Genomes?, 420
Genome-Sequencing Centers, 421
Sequencing Genomes: Strategies, 421
Genomic Sequence Data: From Unfinished to Finished, 422
When Has a Genome Been Fully Sequenced?, 424
Repository for Genome Sequence Data, 425
Genome Annotation: Features of Genomic DNA, 425
Annotation of Genes in Bacteria, 428
Annotation of Genes in Eukaryotes, 428
Summary: Questions from Genome-Sequencing Projects, 429
Perspective, 431
Pitfalls, 431
Discussion Question, 431
Problem, 431
Self-Test Quiz, 432
References, 432

13 **Completed Genomes: Viruses, 437**
Introduction, 437
Classification of Viruses, 438
Bioinformatics Approaches to Problems in Virology, 439
Diversity and Evolution of Viruses, 441
From Phylogeny to Gene Expression: Bioinformatic Approaches to Herpesvirus, 442
Human Immunodeficiency Virus and Need for Bioinformatic Approaches, 446
Bioinformatic Approaches to HIV-1, 452
Using NCBI Resources, 452
Using LANL Database, 453
Bioinformatic Analysis of Viral Genome: Measles Virus, 453
Additional Bioinformatics Resources, 455
Perspectives, 457
Pitfalls, 459
Web Resources, 460
Discussion Questions, 461
Problems, 461
Self-Test Quiz, 461
Suggested Reading, 461
References, 462

14 **Completed Genomes: Bacteria and Archaea, 465**
Introduction, 466
Classification of Bacteria by Morphological Criteria, 466
Classification of Bacteria and Archaea Based on Genome Size and Geometry, 467
Classification of Bacteria and Archaea Based on Lifestyle, 470
Box 14-1. Smallest Genome size: What Is the Minimal Genome Size for Life? 471
Classification of Bacteria Based on Human Disease Relevance, 473

Classification of Bacteria and Archaea Based on Ribosomal RNA Sequences, 474
Classification of Bacteria and Archaea Based on Other Molecular Sequences, 475
Analysis of Prokaryotic Genomes, 478
Nucleotide Composition, 478
Finding Genes, 480
Lateral Gene Transfer, 483
Annotation and Comparison, 486
Comparison of Prokaryotic Genomes, 488
COG, 488
TaxPlot, 489
MUMmer, 490
Perspective, 493
Pitfalls, 495
Web Resources, 496
Discussion Questions, 496
Problems, 496
Self-Test Quiz, 496
Suggested Reading, 497
References, 497

15 **Eukaryotic Genomes: Fungi, 503**
Introduction, 503
Description and Classification of Fungi, 504
Box 15-1. Fungal Taxonomy, 505
Introduction to Budding Yeast *Saccharomyces Cerevisiae*, 505
Sequencing Yeast Genome, 505
Features of Budding Yeast Genome, 506
Exploring Typical Yeast Chromosome, 508
Acquisition of New Genes: Duplication of *S. cerevisiae* Genome, 511
Box 15-2. Gene Nomenclature in *Saccharomyces cerevisiae*, 514
Functional Genomics Projects, 520
Genetic Footprinting Using Transposons, 522
Harnessing Exogenous Transposons, 523
Genomewide Deletions with Molecular Barcodes, 526
Analysis of Fungal Genomes, 528
Aspergillus fumigatus, 529
Candida albicans, 529
Atypical Fungus: Microsporidial Parasite *Encephalitozoon cuniculi*, 529
Neurospora crassa, 529
First Basidiomycete: *Phanerochaete chrysosporium*, 530
Fission Yeast *Schizosaccharomyces pombe*, 530
Perspective, 531
Pitfalls, 532
Web Resources, 533
Discussion Questions, 533
Problems, 534
Self-Test Quiz, 534
Suggested Reading, 535
References, 535

16 **Eukaryotic Genomes: From Parasites to Primates, 539**
Introduction, 539
General Features of Eukaryotes, 541
Major Differences between Eukaryotes and Prokaryotes, 541
C Value Paradox: Why Eukaryotic Genome Sizes Vary So Greatly, 543
Many Eukaryotic Genomes Consist of Noncoding and Repetitive DNA Sequences, 543
Software to Detect Repetitive DNA: RepeatMasker and Censor, 550
Definition of Gene, 551
Finding Genes in Eukaryotic Genomes, 553
Protein-Coding Genes in Eukaryotes: New Paradox, 562
Transcription Factor Databases and Other Genomic DNA Databases, 563
Eukaryotic Genomes Are Organized into Chromosomes, 564

Comparison of Eukaryotic DNA: PipMaker and VISTA, 566
Individual Eukaryotic Genomes, 567
Introduction, 567
Box 16-1. Inconsistent Phylogenies, 570
Protozoans at Base of Tree, 570
Genomes of Unicellular Pathogens: Trypanosomes and *Leishmania*, 571
Malaria Parasite *Plasmodium falciparum* and Other Apicomplexans, 572
Plant Genomes, 575
Overview, 575
Box 16-2. An Astonishing Way to Acquire a Chloroplast: Eat an Organism That Has It, 576
Arabidopsis thaliana Genome, 576
Rice, 579
Box 16-3. Databases for Eukaryotic Genomes, 581
Slime at the Feet of Metazoans, 582
Introduction to Metazoans, 582
Analysis of a Simple Animal: Nematode *Caenorhabditis elegans*, 583
First Insect Genome: *Drosophila melanogaster*, 585
Second Insect Genome: *Anopheles gambiae*, 587
Road to Vertebrates: *Ciona intestinalis*, 587
Vertebrate Genome of a Fish, *Fugu rubripes*, 588
Analysis of Mammalian Genome: Mouse, 589
Primate Genomes, 591
Perspective, 594
Pitfalls, 595
Web Resources, 596
Discussion Questions, 596
Problems, 596
Self-Test Quiz, 597
Suggested Reading, 598
References, 598

17 **Human Genome, 607**
Introduction, 607
Main Conclusions of Human Genome Project, 608
Three Gateways to Access the Human Genome, 609
NCBI, 609
Ensembl, 610
University of California at Santa Cruz Human Genome Browser, 614
Human Genome Project, 617
Background of Human Genome Project, 618
Strategic Issues: Hierarchical Shotgun Sequencing to Generate Draft Sequence, 620
Features of Genome Sequence, 623
Broad Genomic Landscape, 627
Long-Range Variation in GC Content, 627
CpG Islands, 628
Comparison of Genetic and Physical Distance, 629
Repeat Content of Human Genome, 629
Transposon-Derived Repeats, 630
Simple Sequence Repeats, 632
Segmental Duplications, 632
Gene Content of Human Genome, 633
Noncoding RNAs, 635
Protein-Coding Genes, 636
Comparative Proteome Analysis, 636
Complexity of Human Proteome, 637
Perspective, 641
Pitfalls, 641
Web Resources, 642
Discussion Questions, 642
Problem, 642
Self-Test Quiz, 642
Suggested Reading, 643
References, 643

18 **Human Disease, 647**
Human Genetic Disease:
A Consequence of DNA Variation, 647
A Bioinformatics Perspective on Human Disease, 649
Garrod's View of Disease, 650
Categories of Disease, 652
Classification of Disease, 653
NIH Disease Classification: MeSH Terms, 655
Monogenic Disorders, 655
Box 18-1. Sickle Cell Anemia, 657
Box 18-2. Rett Syndrome, 658
OMIM: Central Bioinformatics Resource for Human Disease, 659
Other Central Human Disease Databases, 661
Mutation Databases, 661
Single-Nucleotide Polymorphisms and Disease, 666
Complex Disorders, 669
Box 18-3. Autism: Complex Disorder of Unknown Etiology, 671
Analysis of Chromosomal Abnormalities in Disease, 673
Box 18-4. Genomic Microarrays, 674
Systems-Level Bioinformatics Resources for Human Disease: Organellar and Pathway Databases, 675
Human Disease Genes in Model Organisms, 676
In Mouse, 679
Human Disease Organizations, 684
Functional Classification of Disease Genes, 684
Perspective, 687
Pitfalls, 688
Web Resources, 689
Discussion Questions, 689
Problems, 689
Self-Test Quiz, 690
Suggested Reading, 691
References, 691

Epilogue, 695

Appendix: GCG for Protein and DNA Analysis, 697
Topic 1: Introduction to GCG, 697
Topic 2: General Commands, 697
Topic 3: Entering and Editing Sequences, 699
Topic 4: Pairwise Alignment, 701
Topic 5: Multiple Sequence Alignment, 709
Topic 6: Phylogeny, 712
Topic 7: Sequence Analysis, 713

Glossary, 717

Solutions to Self-Test Quizzes, 735

Subject Index, 737

Author Index, 753

Foreword

Bioinformatics has grown up in the last ten years for one simple reason: data. The increasingly large amount of experimental data has driven the development of data representations and algorithms that are used to manage this information. If the revolution in computation had occurred without a concomitant revolution in biology, computers would still serve biology as they did in the mid-1980s: some scientists (such as structural biologists) would rely on computation for model refinement, simulation and visualization, and a few devoted enthusiasts would lovingly curate special-purpose "boutique" databases. Most biologists would use computers only intermittently–for literature searches, graphing of results, depiction of reigning qualitative models, and perhaps the occasional sequence search or comparison.

However, there *was* a revolution in biology. The ability to determine experimentally the sequence of biological molecules (proteins and nucleic acids) accelerated beyond the expectations of many, and by the early 1990s it was clear that improved methods for storing, analyzing and disseminating biological information would be required. The decision by the scientific community to pursue the human genome project (and associated sequencing efforts) created a market for scientists who understood the challenges in analyzing this data, and possessed skills for creating new methods. The resulting influx of talent from mathematics, statistics, and computer science led to the first generation of breakthroughs that assisted in the assembly of the genome sequences, the creation of the sequence databases, and the development of rapid search and comparison algorithms for sequence analysis.

In its early days, bioinformatics accumulated its practitioners from a variety of ranks. Biologists working on certain types of problems were thrust into computation: x-ray crystallography, NMR spectroscopy, phylogenetics, and population or statistical genetics. At the same time, three-dimensional structural computations attracted the attention of computer scientists with interests in computer graphics and computational geometry. Computer scientists with an interest in algorithms that work on strings found the challenges of sequence analysis enticing. More recently, statisticians have become enamored with the challenges of analyzing noisy but intriguing microarray data. These pioneers created a plethora of methods. Some methods were generally useful, and provided solutions to recurrent problems (e.g. BLAST). Some were successful at solving particular biological problems, but had limited utility to the wider community. Some methods were intriguing demonstrations, but made no biological contribution.

The sequencing projects kicked off the expansion of bioinformatics in two ways. First, it provided the critical mass of challenges that allowed the field to claim a mission, and grow accordingly. The second way is much more fundamental: biologists realized that the fruits of the genome project would be available to them

only through computational analysis. Thus, the field had the critical support of the larger biology community during the early days of its development. This situation led to the development of programs to support fundamental research and training in bioinformatics, and created enough goodwill between biologists and the bioinformaticians to allow the latter group time to deliver a useful set of initial tools, and to create sustainable modes of scientific interaction and productivity.

The future of bioinformatics is clear in some ways, and uncertain in others. There is no doubt that many biologists have embraced the "high throughput" approach to biology, and are cleverly devising new ways to collect fabulously useful information. Following the revolution in sequencing came, in rapid sequence, other technologies for making important measurements on a large scale. The highest impact so far has been the development of microarrays for measuring mRNA expression in cells. The miniaturization of laboratories promises to accelerate our ability to take many other experimental approaches. For example, the analysis of the whole-cell content of proteins (and distinguishing between post-translationally modified versions of proteins) using 2-dimensional gel electropheresis and mass spectroscopy promises a wealth of data. Thus, bioinformatics is secure in the sense that there are no apparent bounds on the ability and interest of biologists to generate data. Indeed, the rate of data accumulation is likely to accelerate, so models will become much more complicated, and will require further innovation in simulation and visualization.

On the other hand, there is some question about the exact modes in which bioinformatics will mature. Some believe that there is an important distinction between "tool builders" and "tool users," where the builders are trained in technical disciplines and generate robust methods in response to the general need articulated by the users. They may even live in specialized "bioinformatics" departments where they interact with other tool builders. The users understand these tools, collaborate in driving their development, but are chiefly interested in them insofar as they help them solve problems. Others believe that bioinformatics will merge into the fabric of biology in a much more integrated manner. The next generation of biologists will have skills and knowledge in quantitative and computational fields, and will be a major source for new representations and algorithms that they will develop in response to problems that they need to solve. The builder/user distinction will not be useful because the very way in which biologists approach problems and solve them will merge bioinformatics with other activities, such as experiment design. As usual, the answer is probably somewhere in the middle. The specialized quantitative and computational knowledge required for tool building will often come from outside biology, but biologists will incorporate these technical disciplines' intellectual and methodological approaches in their own work. This book is an important step in achieving this integration.

In *Bioinformatics and Functional Genomics,* Jonathan Pevsner has created an intriguing text. On the surface, this book is targeted towards biologists who want to solve problems. The organization of the book is based on the fundamental dogma of biology and moves from DNA to RNA to protein, while also moving from single-gene analyses to multiple-gene to genomic. The description of the relevant methods is geared towards helping users understand the key features of the algorithm, while not getting into unnecessary technical detail about implementation.

But there is more to this book. It does not merely describe algorithms that can be used to solve problems in a cookbook fashion. It also provides a very useful

compendium of many major biological insights and breakthroughs over the last decade or so. It cannot be helped: bioinformatics methods are part of the genomic revolution, so in explaining bioinformatics tools, one is forced to describe the biological motivations behind the development of the tool and the biological problems that were solved as a result of their development! Thus, this book becomes a very useful resource not only for biologists (who want to solve problems) but for computational scientists who want to understand the biological problems that bioinformatics has tackled, how well it has solved them, and what outstanding problems remain.

I was repeatedly struck by the depth of biological insight required to properly motivate the creation of tools that were developed and provide meaningful examples. For example, the discussion of molecular phylogeny and evolution (Chapter 11) provides an excellent tutorial on the entire field of molecular evolution—and the description of the computational tools is equally useful! Similarly, the discussion of gene expression (Chapters 6, 7) provides an excellent overview of current models of gene expression, how these data can be used to generate hypotheses, and where it falls short. Finally, the review of the genomes (Chapters 12–17) provides a fantastic overview of the biological challenges facing each type of organism, often justifying its importance as a model organism, and thus clearly defining the outstanding informatics problems that need to be addressed.

Thus, we are left with an exciting new contribution for both the biological community and the computational/informatics community. For the biologists, we have an invaluable guide to the application of non-trivial tools to non-trivial biological problems. The clearly written text is supplemented with perspectives, pitfalls, questions, problems, quizzes and references (online and offline) in order to allow a serious scientist to quickly attain state-of-the-art methods in bioinformatics analysis. For the computational/informatics community, we have a summary of the key problems that these methods are developed to solve, why they are hard, how well current methods perform, and therefore provide a critical introduction to the serious biological motivations for bioinformatics, and a resource for evaluating opportunities and progress in the field. For these reasons, I am enthusiastic about this volume, and wish you the best as you use it both to understand the state of current bioinformatics tools (and how to use them expertly) as well as the biological problems (both solved and outstanding) that motivate them.

Russ Altman
Stanford University

This Page Intentionally Left Blank

Preface

Origins of this Book

This book emerged from lecture notes I prepared several years ago for an introductory bioinformatics and genomics course at the Johns Hopkins School of Medicine. The first class consisted of about 70 graduate students and several hundred auditors, including postdoctoral fellows, technicians, undergraduates, and faculty. Those who attended the course came from a broad variety of fields—students of genetics, neuroscience, immunology or cell biology, clinicians interested in particular diseases, statisticians and computer scientists, virologists and microbiologists. They had a common interest in wanting to understand how they could apply the tools of computer science to solve biological problems. This is the domain of bioinformatics, which I define most simply as the interface of computer science and molecular biology. This emerging field relies on the use of computer algorithms and computer databases to study proteins, genes, and genomes. Functional genomics is the study of gene function using genome-wide experimental and computational approaches.

Comparison

At its essence, the field of bioinformatics is about comparisons. In the first third of the book we learn how to extract DNA or protein sequences from the databases, and then to compare them to each other in a pairwise fashion or by searching an entire database. For the student who has a gene of particular interest, a natural question is to ask "what other genes (or proteins) are related to mine?"

In the middle third of the book, we move from DNA to RNA (gene expression) and to proteins. We again are engaged in a series of comparisons. We compare gene expression in two cell lines with or without drug treatment, or a wildtype mouse heart versus a knockout mouse heart, or a frog at different stages of development. These comparisons extend to the world of proteins, where we apply the tools of proteomics to complex biological samples under assorted physiological conditions. The alignment of multiple, related DNA or protein sequences is another form of comparison. These relationships can be visualized in a phylogenetic tree.

The last third of the book spans the tree of life, and this provides another level of comparison. Which forms of human immunodeficiency virus threaten us, and how can we compare the various HIV subtypes to learn how we might develop a vaccine? How are a mosquito and a fruit fly related? What genes do vertebrates such as fish and humans share in common, and which genes are unique to various phylogenetic lineages?

I believe that these various kinds of comparisons are what distinguish the newly emerging fields of bioinformatics and genomics from traditional biology. Biology has always concerned comparisons; in this book I quote 19th century biologists such as Richard Owen, Ernst Haeckel, and Charles Darwin who engaged in comparative studies at the organismal level. The problems we are trying to solve have not changed substantially. We still seek a more complete understanding of the unifying concepts of biology, such as the organization of life from its constituent parts (e.g., genes and proteins), the behavior of complex biological systems, and the continuity of life through evolution. What *has* changed is how we pursue this more complete understanding. This book describes databases filled with raw information on genes and gene products and the tools that are useful to analyze these data.

The Challenge of Human Disease

My training is as a molecular biologist and neuroscientist. My laboratory studies the molecular basis of childhood brain disorders such as Down syndrome, autism, and lead poisoning. We are located at the Kennedy Krieger Institute, a hospital for children for developmental disorders. (You can learn more about this Institute at http://www.kennedykrieger.org.) Each year over 10,000 patients visit the Institute. The hospital includes clinics for children with a variety of conditions including language disorders, eating disorders, autism, mental retardation, spina bifida, and traumatic brain injury. Some have very common disorders, such as Down syndrome (affecting about 1:700 live births) and mental retardation. Others have rare disorders, such as Rett syndrome or adrenoleukodystrophy.

We are at a time when the number of base pairs of DNA deposited in the world's public repositories has reached tens of billions, as described in Chapter 2. We have obtained the first sequence of the human genome, and since 1995 hundreds of genomes have been sequenced. Throughout the book, you can follow the progress of science as we learn how to sequence DNA, and study its RNA and protein products. At times the pace of progress seems dazzling.

Yet at the same time we understand so little about human disease. For thousands of diseases, a defect in a single gene causes a pathological effect. Even as we discover the genes that are defective in diseases such as cystic fibrosis, muscular dystrophy, adrenoleukodystrophy, and Rett syndrome, the path to finding an effective treatment or cure is obscure. But single gene disorders are not nearly as common as complex diseases such as autism, depression, and mental retardation that are likely due to mutations in multiple genes. And all genetic disease is not nearly as common as infectious disease. We know little about why one strain of virus infects only humans, while another closely related species infects only chimpanzees. We do not understand why one bacterial strain may be pathogenic, while another is harmless. We have not learned how to develop an effective vaccine against any eukaryotic pathogen, from protozoa (such as *Plasmodium falciparum* that causes malaria) to parasitic nematodes.

The prospects for making progress in these areas are very encouraging specifically because of the recent development of new bioinformatics tools. We are only now beginning to position ourselves to understand the genetic basis of both disease-causing agents and the hosts that are susceptible. Our hope is that the information so rapidly accumulating in new bioinformatics databases can be translated through research into insights into human disease and biology in general.

Note to Readers

This book describes over 1,000 websites related to bioinformatics and functional genomics. All of these sites evolve over time (and some become extinct). In an effort to keep the web links up-to-date, a companion website (http://www.bioinfbook.org) maintains essentially all of the website links, organized by chapter of the book. We try our best to maintain this site over time. We use a program to automatically scan all the links each month, and then we update them as necessary.

An additional site is available to instructors, including detailed solutions to problems (see http://www.wiley.com).

Acknowledgments

Writing this book has been a wonderful learning experience. It is a pleasure to thank the many people who have contributed. In particular, the intellectual environment at the Kennedy Krieger Institute and the Johns Hopkins School of Medicine has been extraordinarily rich. These chapters were developed from lectures in an introductory bioinformatics course. The Johns Hopkins faculty who lectured during its first three years were Jef Boeke (yeast functional genomics), Aravinda Chakravarti (human disease), Neil Clarke (protein structure), Kyle Cunningham (yeast), Garry Cutting (human disease), Rachel Green (RNA), Stuart Ray (molecular phylogeny), and Roger Reeves (the human genome). I have benefited greatly from their insights into these areas.

I gratefully acknowledge the many reviewers of this book, including a group of anonymous reviewers who offered extremely constructive and detailed suggestions. Those who read the book include Russ Altman, Christopher Aston, David P. Leader, and Harold Lehmann (various chapters), Conover Talbot (Chapters 2 and 18), Edie Sears (Chapter 3), Tom Downey (Chapter 7), Jef Boeke (Chapter 8 and various other chapters), Michelle Nihei and Daniel Yuan (Chapter 8), Mario Amzel and Ingo Ruczinski (Chapter 9), Stuart Ray (Chapter 11), Marie Hardwick (Chapter 13), Yukari Manabe (Chapter 14), Kyle Cunningham and Forrest Spencer (Chapter 15), and Roger Reeves (Chapter 16). Kirby D. Smith read Chapter 18 and provided insights into most of the other chapters as well. Each of these colleagues offered a great deal of time and effort to help improve the content, and each served as a mentor. Of the many students who read the chapters I mention Rong Mao, Ok-Hee Jeon, and Vinoy Prasad. I particularly thank Mayra Garcia and Larry Frelin who provided invaluable assistance throughout the writing process. I am grateful to my editor at John Wiley & Sons, Luna Han, for her encouragement.

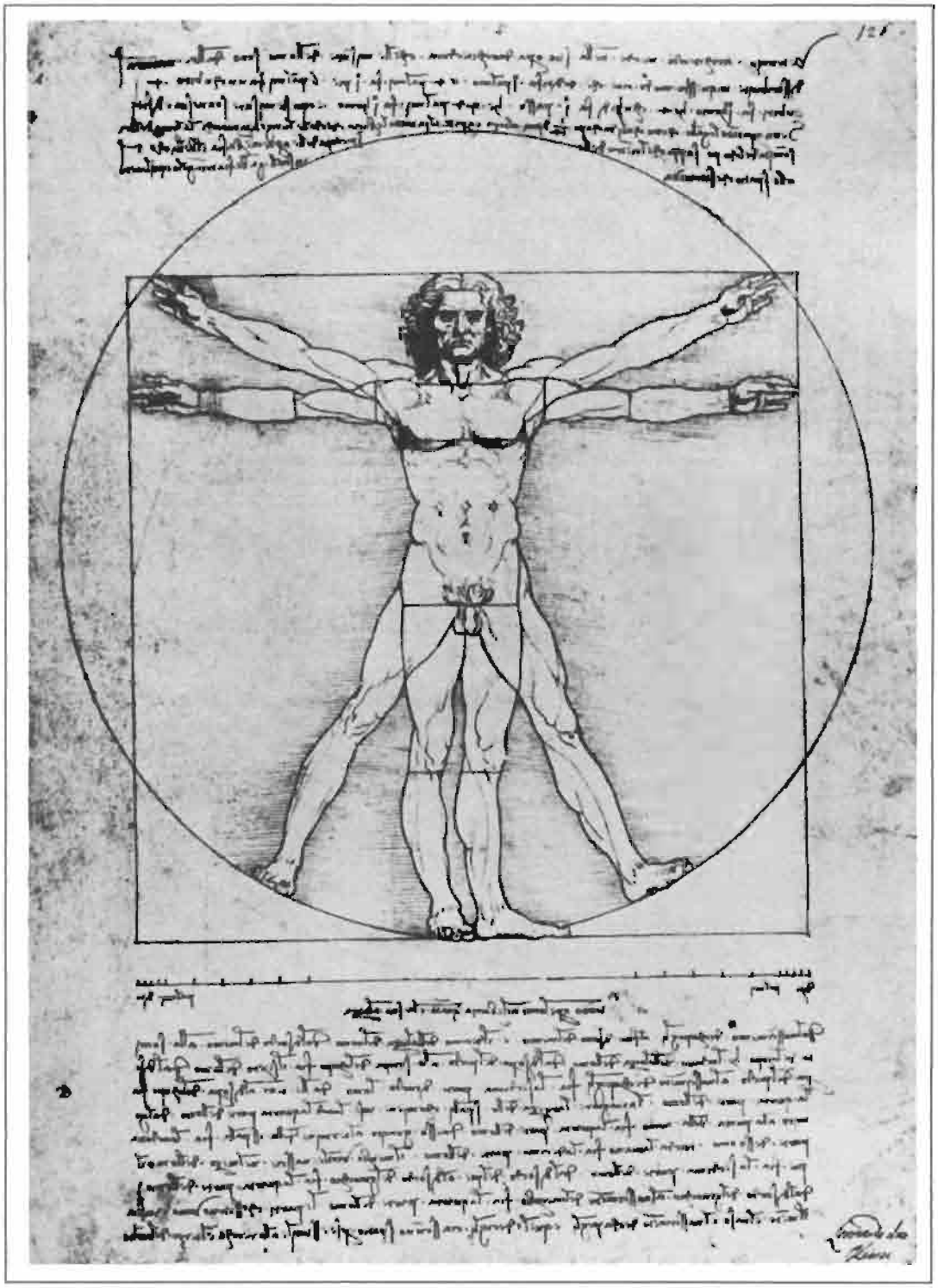
I also acknowledge Gary W. Goldstein, President of the Kennedy Krieger Institute, and Solomon H. Snyder, my chairman in the Department of Neuroscience at Johns Hopkins. Both provided encouragement, and allowed me the opportunity to write this book while maintaining an academic laboratory.

On a personal note, I thank my family for all their love and support, as well as N. Varg, Kimberly Reed, and Charles Cohen. Most of all, I thank my fiancée Barbara Reed for her patience, faith, and love.

This Page Intentionally Left Blank

Part I

Analyzing DNA, RNA, and Protein Sequences in Databases



Leonardo da Vinci (1452–1519) drew the human body in 1490 based on the writings of Vitruvius. This drawing symbolizes Leonardo's quest to unify his art, science and engineering. Leonardo himself is a symbol for the effort to maximize human potential by understanding as many aspects of the human experience as possible. He attempted to study the human body from mathematical principles. The text accompanying this figure reads in part, "If you open the legs so as to reduce the stature by one-fourteenth, and open and raise your arms so that your middle fingers touch the line through the top of the head, know that the center of the extremities of the outspread limbs will be the umbilicus, and the space between the legs will make and equilateral triangle." We can use this image as a symbol to think about the efforts of bioinformatics and genomics to understand all of human biology, from molecular sequences to behavior.

1

Introduction

Bioinformatics represents a new field at the interface of the twentieth-century revolutions in molecular biology and computers. A focus of this new discipline is the use of computer databases and computer algorithms to analyze proteins, genes, and the complete collections of deoxyribonucleic acid (DNA) that comprises an organism (the genome). A major challenge in biology is to make sense of the enormous quantities of sequence data and structural data that are generated by genome-sequencing projects, proteomics, and other large-scale molecular biology efforts. The tools of bioinformatics include computer programs that help to reveal fundamental mechanisms underlying biological problems related to the structure and function of macromolecules, biochemical pathways, disease processes, and evolution.

According to a National Institutes of Health (NIH) definition, bioinformatics is "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, analyze, or visualize such data." The related discipline of computational biology is "the development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems."

The NIH Bioinformatics Definition Committee findings are reported at ▸http://grants.nih.gov/grants/bistic/CompuBioDef.pdf. For additional definitions of bioinformatics and functional genomics, see Boguski (1994), Luscombe et al. (2001), Ideker et al. (2001), and Goodman (2002).

For definitions of functional genomics, see ▸http://bip.weizmann.ac.il/mb/functional_genomics.html.

While the discipline of bioinformatics focuses on the analysis of molecular sequences, genomics and functional genomics are two closely related disciplines. The goal of genomics is to determine and analyze the complete DNA sequence of an organism, that is, its genome. The DNA encodes genes, which can be expressed as ribonucleic acid (RNA) transcripts and then translated into protein. Functional

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

genomics describes the use of genomewide assays to the study of gene and protein function.

The aim of this book is to explain both the theory and practice of bioinformatics. The book is especially designed to help the biology student use computer programs and databases to solve biological problems related to proteins, genes, and genomes. Bioinformatics is an integrative discipline, and our focus on individual proteins and genes is part of a larger effort to understand broad issues in biology such as the relationship of structure to function, development, and disease.

Organization of The Book

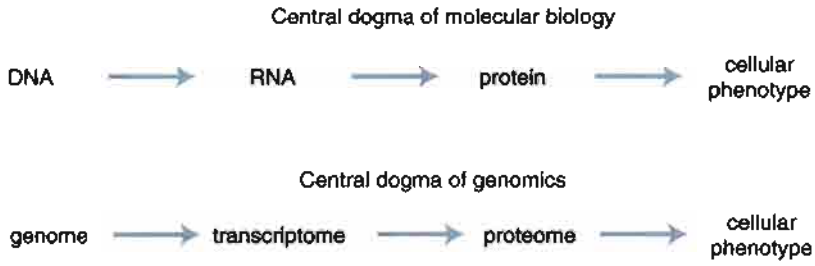
There are three main sections of the book. The first part explains how to access biological sequence data, particularly DNA and protein sequences (Chapter 2). Once sequences are obtained, we show how to compare two sequences (pairwise alignment; Chapter 3) and how to compare multiple sequences [primarily by the Basic Local Alignment Search Tool (BLAST); Chapters 4 and 5].

The second part of the book describes functional genomics approaches to RNA and protein. The central dogma of biology states that DNA is transcribed into RNA then translated into protein. We will examine gene expression, including a description of the emerging technology of DNA microarrays (Chapters 6 and 7). We then consider proteins from the perspective of protein families, the analysis of individual proteins, protein structure, and multiple sequence alignment (Chapters 8–10). The relationships of protein and DNA sequences that are multiply aligned can be visualized in phylogenetic trees (Chapter 11). Chapter 11 thus introduces the subject of molecular evolution.

Since 1995, the genomes have been sequenced for several hundred bacteria and archaea as well as fungi, animals, and plants. The third section of the book covers genome analysis. Chapter 12 provides an overview of the study of completed genomes and then descriptions of how the tools of bioinformatics can elucidate the tree of life. We describe bioinformatics resources for the study of viruses (Chapter 13) and bacteria and archaea (Chapter 14; these are two of the three main branches of life). Next we examine a variety of eukaryotes (from fungi to primates; Chapters 15 and 16) and then the human genome (Chapter 17). Finally, we explore bioinformatic approaches to human disease (Chapter 18).

Bioinformatics: The Big Picture

We can summarize the entire field of bioinformatics with three perspectives. The first perspective on bioinformatics is the cell (Fig. 1.1). The central dogma of molecular biology is that DNA is transcribed into RNA and translated into protein. The focus of molecular biology has been on individual genes, messenger RNA (mRNA) transcripts, and proteins. A focus of the field of bioinformatics is the complete collection of DNA (the genome), RNA (the transcriptome), and protein sequences (the proteome) that have been amassed (Henikoff, 2002). These millions of molecular sequences present both great opportunities and great challenges. A bioinformatics approach to molecular sequence data involves the application of computer algorithms and computer databases to molecular and cellular biology. Such an approach is sometimes referred to as functional genomics. This typifies the essential nature of bioinformatics: biological questions can be approached from levels ranging from



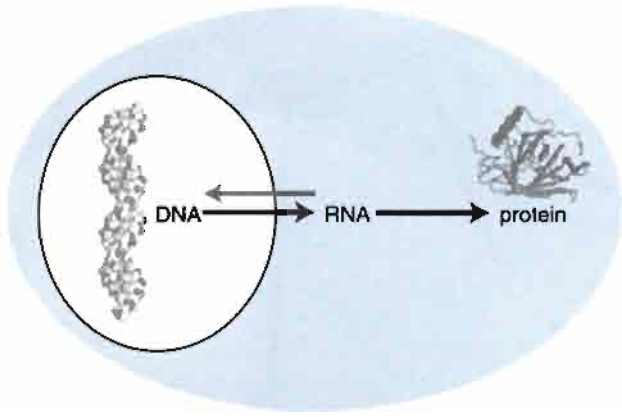


FIGURE 1.1. The first perspective of the field of bioinformatics is the cell. Bioinformatics has emerged as a discipline as biology has become transformed by the emergence of molecular sequence data. Databases such as the European Molecular Biology Laboratory (EMBL), GenBank, and the DNA Database of Japan (DDBJ) serve as repositories for billions of nucleotides of DNA sequence data (see Chapter 2). Corresponding databases of expressed genes (RNA) and protein have been established. A main focus of the field of bioinformatics is to study molecular sequence data to gain insight into a broad range of biological problems.

single genes and proteins to cellular pathways and networks or even whole genomic responses (Ideker et al., 2001). Our goals are to understand how to study both individual genes and proteins and collections of thousands of genes/proteins.

From the cell we can focus on individual organisms, which represents the second perspective of the field of bioinformatics (Fig. 1.2). Each organism changes across different stages of development and (for multicellular organisms) across different regions of the body. For example, while we may sometimes think of genes

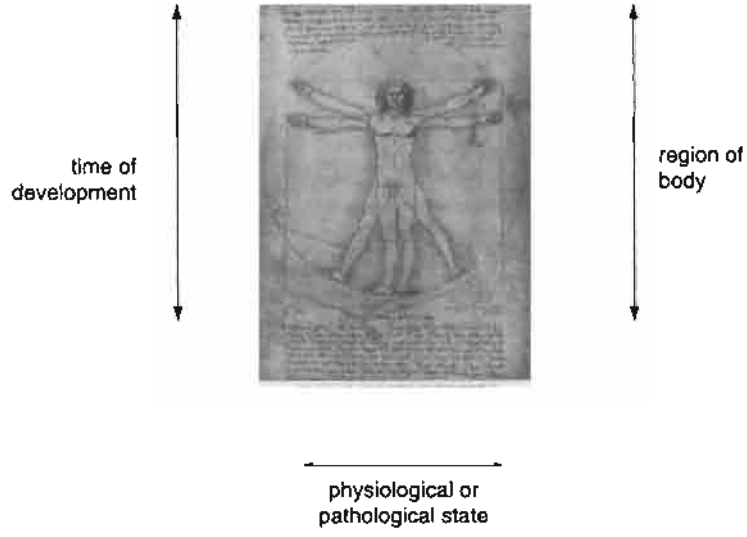


FIGURE 1.2. The second perspective of bioinformatics is the organism. Broadening our view from the level of the cell to the organism, we can consider the individual's genome (collection of genes), including the genes that are expressed as RNA transcripts and the protein products. Thus, for an individual organism bioinformatics tools can be applied to describe changes through developmental time, changes across body regions, and changes in a variety of physiological or pathological states.

FIGURE 1.3. The third perspective of the field of bioinformatics is represented by the tree of life. The scope of bioinformatics includes all of life on Earth, including the three major branches of bacteria, archaea, and eukaryotes. Viruses, which exist on the borderline of the definition of life, are not depicted here. For all species, the collection and analysis of molecular sequence data allow us to describe the complete collection of DNA that comprises each organism (the genome). We can further learn the variations that occur between species and among members of a species, and we can deduce the evolutionary history of life on Earth. (After Pace, 1997.) Used with permission.

as static entities that specify features such as eye color or height, they are in fact dynamically regulated across time and region and in response to physiological state. Gene expression varies in disease states or in response to a variety of signals, both intrinsic and environmental. Many bioinformatics tools are available to study the broad biological questions relevant to the individual: There are many databases of expressed genes and proteins derived from different tissues and conditions. One of the most powerful applications of functional genomics is the use of DNA microarrays to measure the expression of thousands of genes in biological samples.

At the largest scale is the tree of life (Fig. 1.3) (Chapter 12). There are many millions of species alive today, and they can be grouped into the three major branches of bacteria, archaea (single-celled microbes that tend to live in extreme environments), and eukaryotes. Molecular sequence databases currently hold DNA sequence from

over 100,000 different organisms. The complete genome sequences of several hundred organisms will soon become available. One of the main lessons we are learning is the fundamental unity of life at the molecular level. We are also coming to appreciate the power of comparative genomics, in which genomes are compared.

Figure 1.4 on the following page presents the contents of this book in the context of the three perspectives of bioinformatics.

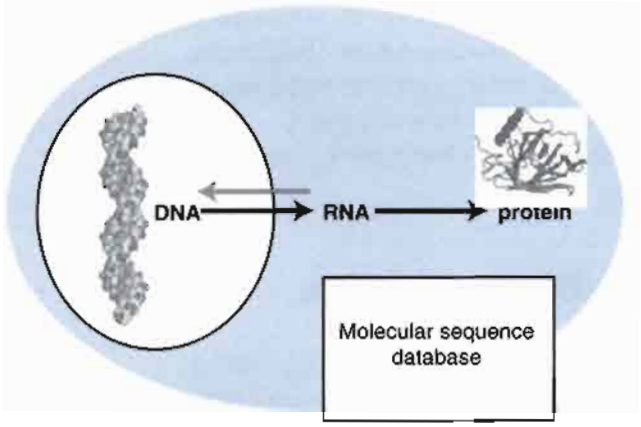
A Consistent Example: Retinol-Binding Protein

Throughout this book we will focus on the example of a gene and its corresponding protein product: retinol-binding protein (RBP4), a small, abundant secreted protein that binds retinol (vitamin A) in blood (Newcomer and Ong, 2000). Retinol, obtained from carrots in the form of vitamin A, is very hydrophobic. RBP4 helps transport this ligand to the eye where it is used for vision. We will study RBP4 in detail because it has a number of interesting features:

- There are many proteins that are homologous to RBP4 in a variety of species, including human, mouse, and fish ("orthologs"). We will use these as examples of how to align proteins, perform database searches, and study phylogeny (Chapters 2–11).
- There are other human proteins that are closely related to RBP4 ("paralogs"). Altogether the family that includes RBP4 is called the lipocalins, a diverse group of small ligand-binding proteins that tend to be secreted into extracellular spaces (Akerstrom et al., 2000; Flower et al., 2000). Other lipocalins have fascinating functions such as apoliprotein D (which binds cholesterol), a pregnancy-associated lipocalin, aphrodisin (an "aphrodisiac" in hamsters), and an odorant-binding protein in mucus.
- There are even bacterial lipocalins, which could have a role in antibiotic resistance (Bishop, 2000). We will explore how bacterial lipocalins could be ancient genes that entered eukaryotic genomes by a process called lateral gene transfer.
- The gene expression levels of some lipocalins are dramatically regulated (Chapters 6 and 7).
- Because the lipocalins are small, abundant, and soluble proteins, their biochemical properties have been characterized in detail. The three-dimensional protein structure has been solved for several of them by X-ray crystallography (Chapter 9).
- Some lipocalins have been implicated in human disease (Chapter 18).

Another molecule we will introduce is the *pol* (polymerase) gene of human immunodeficiency virus 1 (HIV-1). HIV presents one of the greatest public health challenges in the world today. Over 42 million people are infected as of the end of the year 2002 and over 16 million people have died. The HIV-1 genome encodes just nine proteins, including *pol* (Frankel and Young, 1998). We will examine *pol* throughout the book because the properties of this gene, its protein products, and the HIV-1 genome are distinct from the lipocalins.

- The *pol* gene is a multidomain protein: it is a single polypeptide with several structurally and functionally distinct domains. The *pol* gene encodes a protein of 1003 amino acids with reverse transcriptase activity (that is, an

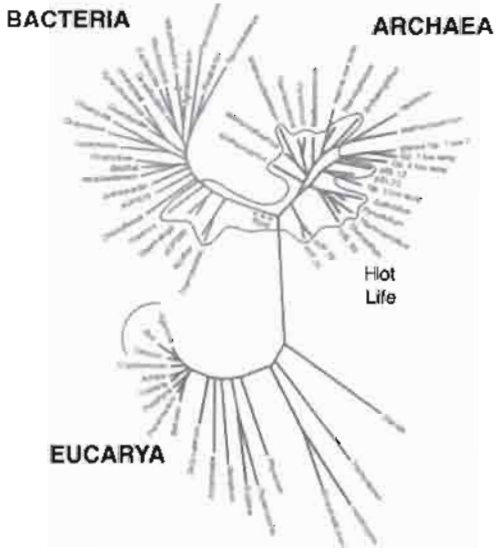


Part 1: Analyzing DNA, RNA, and protein sequences

Chapter 1: Introduction
Chapter 2: How to obtain sequences
Chapter 3: How to compare two sequences
Chapters 4 and 5: How to compare a sequence to all other sequences in databases

Part 2: Genome-wide analysis of RNA and protein

Chapter 6: Gene expression
Chapter 7: Microarrays
Chapter 8: Protein analysis and protein families
Chapter 9: Protein structure
Chapter 10: How to multiply align sequences
Chapter 11: How to view multiply aligned sequences as phylogenetic trees



Part 3: Genome analysis

Chapter 12: The tree of life
Chapter 13: Viruses
Chapter 14: Prokaryotes
Chapters 15 and 16: Eukaryotes
Chapter 17: The human genome
Chapter 18: Human disease

FIGURE 1.4. Overview of the chapters in this book.

RNA-dependent DNA polymerase). It is also an aspartyl protease, and it has integrase activity. These multiple activities are typical of multidomain proteins.

- The modular nature of the pol protein affects our ability to perform database searches (Chapters 4 and 5) and multiple sequence alignments (Chapters 8 and 10).
- The *pol* gene incorporates substitutions extremely rapidly. A typical individual infected by HIV may have over a million variants of *pol*. The study of the evolution of *pol* complements our study of the lipocalins (Chapter 11).
- As a viral protein, our study of pol gives us the opportunity to learn how to access bioinformatics resources relevant to studying viruses (Chapter 13). Database searches with pol will help emphasize how to restrict searches to particular domains of the tree of life.

Organization of The Chapters

The chapters of this book are intended to provide both the theory of bioinformatics subjects as well as a practical guide to using computer databases and algorithms. Web resources are provided throughout each chapter. Chapters end with brief sections called Perspective and Pitfalls. The perspective feature describes the rate of growth of the subject matter in each chapter. For example, a perspective on Chapter 2 (access to sequence information) is that the amount of DNA sequence data deposited in GenBank is undergoing an explosive rate of growth. In contrast, an area such as pairwise sequence alignment, which is fundamental to the entire field of bioinformatics (Chapter 3), was firmly established in the 1970s and 1980s.

The website for this book (▸http://www.bioinfbook.org) contains about 1000 URLs, organized by chapter.)

The pitfalls section of each chapter describes some common difficulties encountered by biologists using bioinformatics tools. Some errors might seem trivial, such as searching a DNA database with a protein sequence. Other pitfalls are more subtle, such as artifacts caused by multiple sequence alignment programs depending upon the type of algorithm that is selected. Indeed, while the field of bioinformatics depends substantially on analyzing sequence data, it is important to recognize that there are many categories of errors associated with data generation, collection, storage, and analysis.

Each chapter offers multiple-choice quizzes, which test your understanding of the chapter materials. There are also problems that require you to apply the concepts presented in each chapter. These problems may form the basis of a computer laboratory for a bioinformatics course.

The references at the end of each chapter are accompanied by an annotated list of recommended articles. This suggested reading section includes classic papers that show how the principles described in each chapter were discovered. Particularly helpful review articles and research papers are highlighted.

Suggestions For Students and Teachers: Web Exercises and Find-a-Gene

Often, students of bioinformatics have a particular research area of interest such as a gene, a physiological process, a disease, or a genome. It is hoped that by studying RBP4 and other specific proteins and genes throughout this book, students

can simultaneously apply the principles of bioinformatics to their own research questions.

In teaching a course on bioinformatics at Johns Hopkins, it has been helpful to complement lectures with computer labs. All the websites described in this book are freely available on the World Wide Web, and many of the software packages are free for academic use.

Another feature of the Johns Hopkins course is that each student is required to discover a novel gene by the last day of the course. The student must begin with any protein sequence of interest and perform database searches to identify genomic DNA that encodes a protein no one has described before. This problem is described in Chapter 5 (see Fig. 5.17). The student thus chooses the name of the gene and its corresponding protein and describes information about the organism and evidence that the gene has not been described before. Then, the student creates a multiple sequence alignment of the new protein (or gene) and creates a phylogenetic tree showing its relation to other known sequences.

Each year, some beginning students are slightly apprehensive about accomplishing this exercise, but in the end all of them succeed. A benefit of this exercise is that it requires a student to actively use the principles of bioinformatics. Most students choose a gene (or protein) relevant to their own research area, while others find new lipocalins.

Teaching bioinformatics is notable for the diversity of students learning this new discipline. Each chapter provides background on the subject matter. For more advanced students, several key research papers are cited at the end of each chapter. These papers are technical, and reading them along with the chapters will provide a deeper understanding of the material. The suggested reading section also includes review articles.

KEY BIOINFORMATICS WEBSITES

The field of bioinformatics relies heavily on the Internet as a place to access sequence data, to access software that is useful to analyze molecular data, and as a place to integrate different kinds of resources and information relevant to biology. We will describe a variety of websites. Initially, we will focus on the three main publicly accessible databases that serve as repositories for DNA and protein data (Table 1.1). In Chapter 2 we begin with the National Center for Biotechnology

TABLE 1-1 Three Primary Bioinformatics Web Servers That Serve as Centralized Repositories for DNA and Protein Sequence Data

These will be introduced in Chapter 2

Resource	Description	URL[a]
DNA Data Bank of Japan (DDBJ)	Associated with the Center for Information Biology	▸http://www.ddbj.nig.ac.jp/
European Bioinformatics Institute (EBI)	Maintains the EMBL database	▸http://www.ebi.ac.uk/
National Center for Biotechnology Information (NCBI)	Maintains GenBank	▸http://www.ncbi.nlm.nih.gov/

[a] Uniform Resource Locator.

TABLE 1-2 Additional Bioinformatics Web Servers

Each website contains access to dozens of software tools, research projects, literature references, and other information relevant to bioinformatics

Resource	Description	URL
Centre for Molecular and Biomolecular Informatics	From the University of Nijmegen	http://www.cmbi.kun.nl/
ExPASy (Expert Protein Analysis System)	Proteomics server of the Swiss Institute of Bioinformatics	http://www.expasy.org/
GENESTREAM	Institut de Génétique Humaine, Montpellier	http://www2.igh.cnrs.fr/
GenomeNet	In Kyoto	http://www.genome.ad.jp/
INFOBIOGEN	In Montpellier	http://www.infobiogen.fr/page_accueil_en.html
Oak Ridge National Laboratory (ORNL)	In Tennessee	http://compbio.ornl.gov/
Protein Information Resource (PIR)	A Division of National Biomedical Research Foundation	http://pir.georgetown.edu/
The Wellcome Trust Sanger Institute	A genome research center in Cambridge	http://www.sanger.ac.uk/
The Institute for Genomic Research (TIGR)	In Rockville, Maryland	http://www.tigr.org/

Information (NCBI), which hosts GenBank. The NCBI website offers a variety of other bioinformatics-related tools. We will gradually introduce the European Bioinformatics Institute (EBI) web server, which hosts a complementary DNA database (EMBL, the European Molecular Biology Laboratory database). We will also introduce the DNA Database of Japan (DDBJ). The research teams at GenBank, EMBL, and DDBJ share sequence data on a daily basis. A general theme of the discipline of bioinformatics is that many databases are closely interconnected.

Throughout the chapters of this book we will introduce several hundred additional websites that are relevant to bioinformatics. Table 1.2 lists several additional servers that offer databases as well as many programs for the analysis of biological sequences. Table 1.3 lists several additional sites that offer links to bioinformatics resources. We present them now for those who wish to explore the types of bioinformatics resources that are currently available.

TABLE 1-3 Bioinformatics Sites with Useful Links

Websites that provide links to bioinformatics resources

Resource	Description	URL
Amos' WWW links page	From ExPASy	http://www.expasy.ch/alinks.html
DBCAT, The Public Catalog of Databases	From INFOBIOGEN	http://www.infobiogen.fr/services/dbcat/
European Molecular Biology Network	Various European nodes	http://www.embnet.org/
Human Genome Most Used Links	From Los Alamos National Laboratories	http://www-ls.lanl.gov/HGhotlist.html

SUGGESTED READING

Overviews of the field of bioinformatics have been written by Mark Gerstein and colleagues (Luscombe et al., 2001) and Claverie et al. (2001). Kaminski (2000) also introduces bioinformatics, with practical suggestions of websites to visit. Russ Altman (1998) discusses the relevance of bioinformatics to medicine, while David Searls (2000) introduces bioinformatics tools for the study of genomes.

REFERENCES

Akerstrom, B., Flower, D. R., and Salier, J. P. Lipocalins: Unity in diversity. *Biochim. Biophys. Acta* **1482,** 1-8 (2000).

Altman, R. B. Bioinformatics in support of molecular medicine. *Proc. AMIA Symp.,* 53-61 (1998).

Bishop, R. E. The bacterial lipocalins. *Biochim. Biophys. Acta* **1482,** 73-83 (2000).

Boguski, M. S. Bioinformatics. *Curr. Opin. Genet. Dev.* **4,** 383-388 (1994).

Claverie, J. M., Abergel, C., Audic, S., and Ogata, H. Recent advances in computational genomics. *Pharmacogenomics* **2,** 361-372 (2001).

Flower, D. R., North, A. C., and Sansom, C. E. The lipocalin protein family: Structural and sequence overview. *Biochim. Biophys. Acta* **1482,** 9-24 (2000).

Frankel, A. D., and Young, J. A. HIV-1: Fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67,** 1-25 (1998).

Goodman, N. Biological data becomes computer literate: New advances in bioinformatics. *Curr. Opin. Biotechnol.* **13,** 68-71 (2002).

Henikoff, S. Beyond the central dogma. *Bioinformatics* **18,** 223-225 (2002).

Ideker, T., Galitski, T., and Hood, L. A new approach to decoding life: Systems biology. *Annu. Rev. Genomics Hum. Genet.* **2,** 343-372 (2001).

Kaminski, N. Bioinformatics. A user's perspective. *Am J. Respir. Cell Mol. Biol.* **23,** 705-711 (2000).

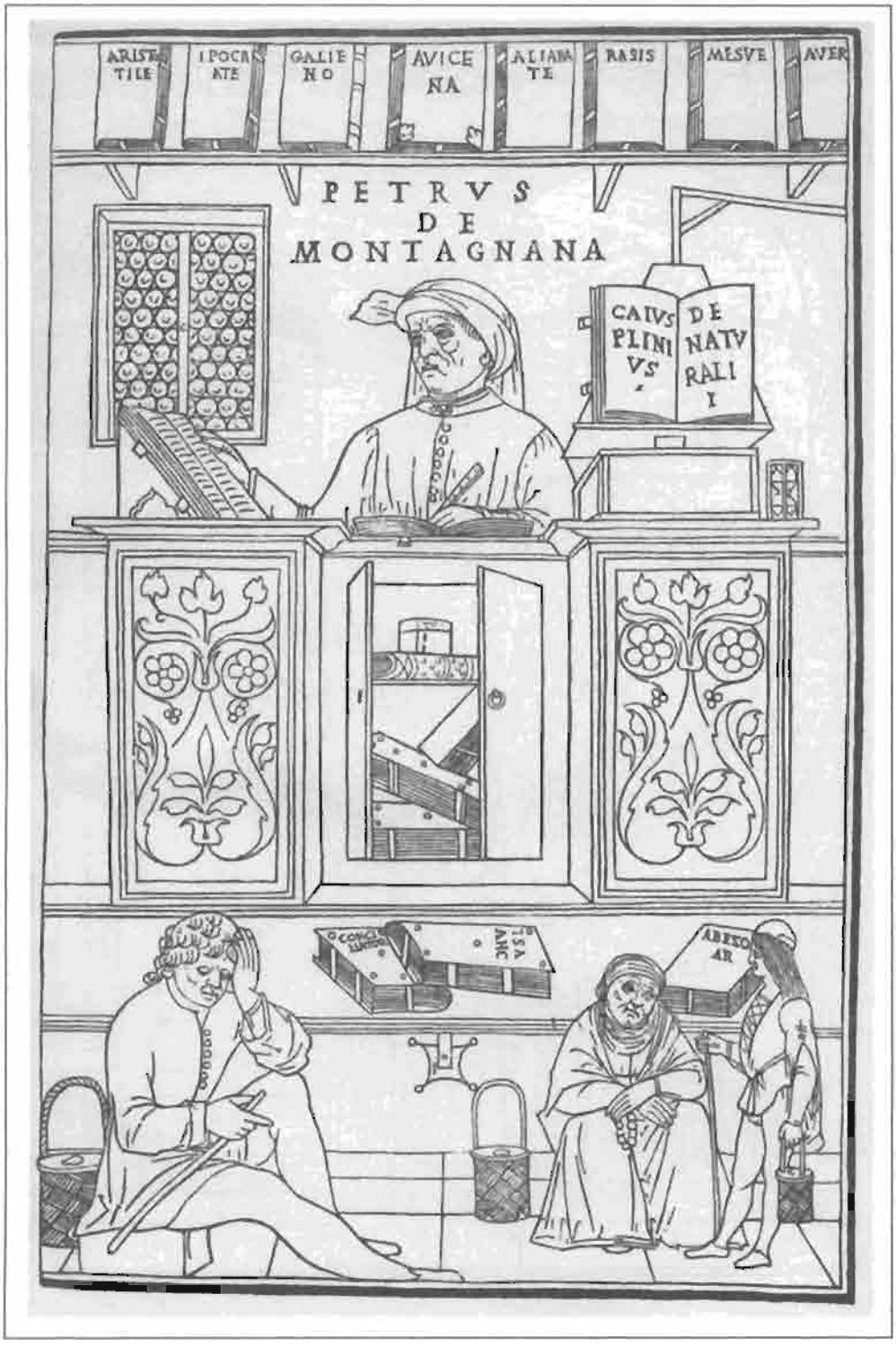
Luscombe, N. M., Greenbaum, D., and Gerstein, M. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med.* **40,** 346-358 (2001).

Newcomer, M. E., and Ong, D. E. Plasma retinol binding protein: Structure and function of the prototypic lipocalin. *Biochim. Biophys. Acta* **1482,** 57-64 (2000).

Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276,** 734-740 (1997).

Searls, D. B. Bioinformatics tools for whole genomes. *Annu. Rev. Genomics Hum. Genet.* **1,** 251-279 (2000).

This Page Intentionally Left Blank



Chapter 2 introduces ways to access molecular data. This image is from the Fasciculus medicinae *of Johannes de Ketham, first published in 1491. The scene shows three patients sitting at the office of Petrus de Montagnana, a doctor and professor at the medical school in Padua (Singer, 1925 depicting a plate from the 1493 edition). A variety of books are at his disposal. Pliny's* Natural History *lies open on a stand on the desk. On the shelf above are books by Aristotle, Hippocrates, Galen, Ibn Sina, Haly Abbas, Rhazes, Mesue, and Averroes. On the floor are works by Pietro d'Abano (the* Conciliator*), Isaac, and Avenzoar. Johannes Gutenberg invented his printing press in the 1450s, and by the year 1500 over two million volumes had been printed in the city of Venice alone (Sarton, 1955). In a sense, during our current genomic era (from the 1950s to the present) we are experiencing a comparable rise in our ability to access information.*

2

Access to Sequence Data and Literature Information

Introduction to Biological Databases

All living organisms are characterized by the capacity to reproduce and evolve. The genome of an organism is defined as the collection of DNA within that organism, including the set of genes that encode proteins. In 1995 the complete genome of a free-living organism was sequenced for the first time, the bacterium *Haemophilus influenzae* (Fleischmann et al., 1995; Chapters 12 and 14). In the few years since then the genomes of dozens of organisms have been completely sequenced, ushering in a new era of biological data acquisition and information accessibility. Publicly available databanks now contain billions of nucleotides of DNA sequence data collected from thousands of different organisms. The goal of this chapter is to introduce the databases that store these data and strategies to extract information from them.

Three publicly accessible databases store large amounts of nucleotide and protein sequence data (Table 1.1): GenBank at the National Center for Biotechnology Information (NCBI) of the National Institutes of Health (NIH) in Bethesda, the DNA Database of Japan (DDBJ), and the European Bioinformatics Institute (EBI) in Hinxton, England. These three databases will be described in this chapter.

These three databases share their sequence data daily as part of the International Nucleotide Sequence Database Collaboration.

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

In addition to GenBank, DDBJ, and EBI, there are other categories of bioinformatics databases that contain DNA and/or protein sequence data:

- Databases such as Ensembl, NCBI, and the Golden Path at the University of California, Santa Cruz (UCSC) provide annotation of the human genome and other genomes (see Chapters 16–17).
- Some contain nucleotide and/or protein sequence data that are relevant to a particular gene or protein (such as kinases). Other databases are specific to particular chromosomes or organelles (Chapters 16–18).
- A variety of databases include information on sequences sharing common properties that have been grouped together. For example, the Protein Family (Pfam) database consists of several thousand families of homologous proteins.

Pfam (▸ http://www.sanger.ac.uk/Software/Pfam/) and other related databases will be described in Chapter 8 (protein families).

- Hundreds of databases contain sequence information related to genes that are mutated in human disease. These databases are described in Chapter 18.
- Many specialized databases focus on particular organisms (such as yeast); examples are listed in the section on genomes (Chapters 12–17).
- There are databases devoted to particular types of nucleic acids or proteins or properties of these macromolecules. Examples are databases of gene expression (see Chapters 6 and 7), databases of transfer RNA (tRNA) molecules, databases of tissue-specific protein expression (see Chapter 8), or databases of gene regulatory regions such as 3′ untranslated regions (see Chapter 16).
- Some databases contain nucleotide and protein sequence data that are not publicly available. Companies such as Celera Genomics offer subscription-based access to these databases (see Chapter 17 on the human genome).

  Some bioinformatics databases do not contain nucleotide or protein sequence data as their main function. Instead, they contain information that may link to individual genes/proteins.
- Literature databases contain bibliographic references relevant to biological research and in some cases contain links to full-length articles. We will describe two of these databases, PubMed and the Sequence Retrieval System (SRS), in this chapter.
- Structure databases contain information on the structure of proteins and other macromolecules. These databases will be described in Chapter 8 (on proteins) and Chapter 9 (on protein structure).

GenBank: Database of Most Known Nucleotide and Protein Sequences

In Chapter 5, we will use specialized searches of databases from assorted genome-sequencing projects. Much of the sequence data from these sequencing projects is publicly available but not yet entered in GenBank.

While the sequence information underlying DDBJ, EBI, and GenBank are equivalent, we begin our discussion with GenBank. GenBank is a database consisting of most known public DNA and protein sequences (Benson et al., 2002). In addition to storing these sequences, GenBank contains bibliographic and biological annotation. Data from GenBank are available free of charge from the National Center for Biotechnology Information (NCBI) in the National Library of Medicine at the NIH (Benson et al., 2002).

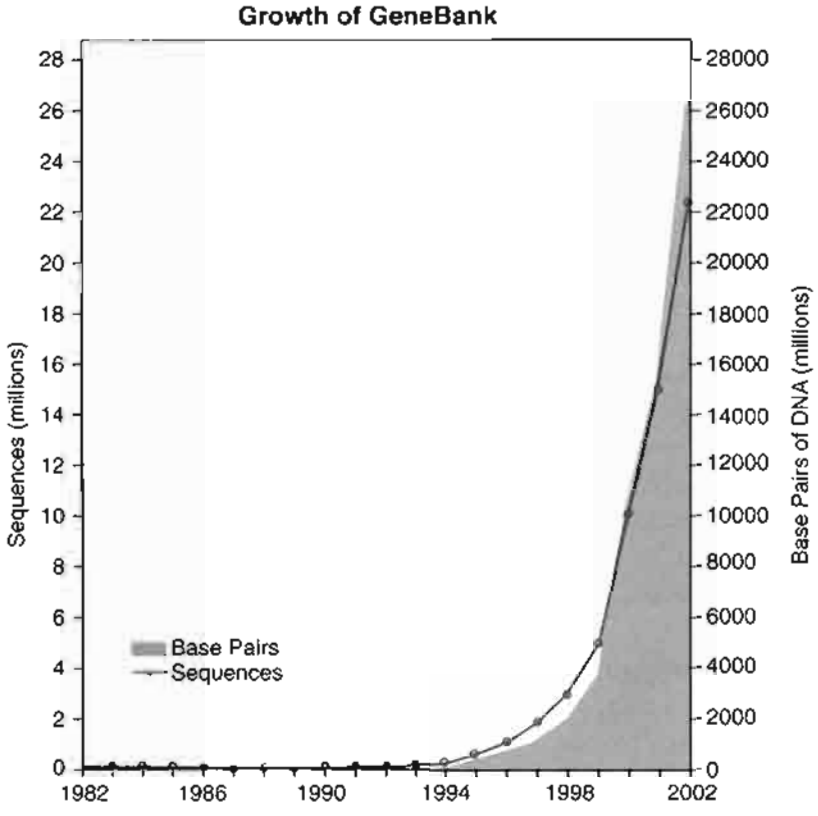


FIGURE 2.1. Growth of GenBank (from ▸http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html).

Amount of Sequence Data

GenBank contains over 31 billion nucleotides from 24 million sequences (release 135.0, April 2003). The growth of GenBank in terms of both nucleotides of DNA and number of sequences from 1982 to 2000 is summarized in Figure 2.1. Over the period 1982 to the present, the number of bases in GenBank has doubled approximately every 14 months.

Between October 2001 and October 2002, over 12 billion base pairs (bp) of DNA were added to GenBank, an average of 33 million bp per day. The first eukaryotic genome to be completed (*Saccharomyces cerevisiae*; Chapter 15) is about 12 million bp in size.

Organisms in GenBank

Over 100,000 different species are represented in GenBank, with over 1000 new species added per month (Benson et al., 2002). The number of organisms represented in GenBank is shown in Table 2.1. We will define the bacteria, archaea, and eukaryotes in detail in Chapters 14–16. Briefly, eukaryotes have a nucleus and are

You can download all of the sequence data in GenBank at the website ▸ftp://ftp.ncbi.nih.gov/genbank. For release 132.0 in October 2002, the total size of these files is about 83 gigabytes (83×10^9 bytes). By comparison, all the words in the United States Library of Congress add up to 20 terabytes (20×10^{12} bytes; 20 trillion bytes). And the particle accelerator used by physicists at CERN near Geneva (▸http://public.web.cern.ch/Public/) collects one petabyte of data each year (10^{15} bytes; 1 quadrillion bytes).

TABLE 2-1 Species Represented in GenBank (1995–2001)

	1995	1996	1997	1998	1999	2000	2001
All species	15,925	22,862	32,569	43,159	61,525	87,168	113,940
Viruses	1,740	1,964	2,513	2,794	3,401	4,165	5,436
Bacteria	2,899	3,798	6,015	8,625	14,209	22,616	28,186
Archaea	154	227	376	544	1,003	1,697	2,038
Eukaryota	10,339	15,863	22,539	29,844	41,295	56,800	76,099

Source: From ▸http://www.ncbi.nlm.nih.gov/Taxonomy/.

TABLE 2-2 Twenty Most Sequenced Organisms in GenBank (Release 135.0, April 2003)

Entries	Bases	Species	Common Name
6,574,171	9,743,398,611	*Homo sapiens*	Human
4,839,873	5,815,119,777	*Mus musculus*	Mouse
702,180	5,565,107,808	*Rattus norvegicus*	Rat
481,072	725,258,089	*Danio rerio*	Zebrafish
350,741	695,203,285	*Drosophila melanogaster*	Fruit fly
221,048	620,437,677	*Oryza sativa (japonica cultivar-group)*	Rice
730,719	402,816,176	*Zea mays*	Corn
502,590	394,829,161	*Arabidopsis thaliana*	Thale cress
567,863	386,355,126	*Brassica oleracea*	Broccoli
13,379	326,808,471	*Macaca mulatta*	Rhesus monkey
427,323	317,823,422	*Gallus gallus*	Chicken
499,207	294,135,749	*Ciona intestinalis*	Ascidian
390,231	223,252,029	*Bos taurus*	Cow
199,302	222,094,921	*Caenorhabditis elegans*	Worm
418,485	207,706,076	*Triticum aestivum*	Wheat
166,243	191,992,299	*Pan troglodytes*	Chimpanzee
189,208	170,811,939	*Tetraodon nigroviridis*	Pufferfish
273,621	166,171,916	*Xenopus laevis*	Frog
282,062	156,196,664	*Hordeum vulgare subsp. vulgare*	Barley
186,121	154,322,129	*Medicago truncatula*	Legume

Source: From ▸ ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt

often multicellular, while bacteria do not have a nucleus. Archaea are single-celled organisms, distinct from eukaryotes and bacteria, which constitute a third major branch of life. Viruses, which contain nucleic acids (DNA or RNA) but can only replicate in a host cell, exist at the borderline of the definition of living organisms.

We have seen so far that GenBank is very large and growing rapidly. From Table 2.1 we see that the organisms in GenBank consist mostly of eukaryotes. Of the microbes, there are currently over 10 times more bacteria than archaea represented in GenBank.

We will discuss how genomes of various organisms are selected for complete sequencing in Chapter 12.

The number of entries and bases of DNA/RNA for the 20 most sequenced organisms in GenBank is provided in Table 2.2 (excluding chloroplast and mitochondrial sequences). This list includes some of the most common model organisms that are studied in biology. Notably, the scientific community is studying a series of mammals (e.g., human, mouse, cow), invertebrates (fruit fly, worm), plants (broccoli, corn, rice), and parasites (African trypanosome and amoebae). Different species are useful for a variety of different studies. Bacteria, archaea, and viruses are absent from the list in Table 2.2 because they have relatively small genomes.

The International Human Genome Sequencing Consortium adopted the Bermuda Principles in 1996, calling for the rapid release of raw genomic sequence data. You can read about recent versions of these principles at ▸ http://www.genome.gov/page.cfm?page ID-10506537.

To help organize the available information, each sequence name in a GenBank record is followed by its data file division and primary accession number. (We will define accession numbers below.) The following codes are used to designate the data file divisions:

1. PRI: primate sequences
2. ROD: rodent sequences

3. MAM: other mammalian sequences
4. VRT: other vertebrate sequences
5. INV: invertebrate sequences
6. PLN: plant, fungal, and algal sequences
7. BCT: bacterial sequences
8. VRL: viral sequences
9. PHG: bacteriophage sequences
10. SYN: synthetic sequences
11. UNA: unannotated sequences
12. EST: EST sequences (expressed sequence tags)
13. PAT: patent sequences
14. STS: STS sequences (sequence-tagged sites)
15. GSS: GSS sequences (genome survey sequences)
16. HTG: HTGS sequences (high-throughput genomic sequences)

We will define the terms STS, GSS, EST and HTGS below.

Types of Data in GenBank

There is an enormous number of molecular sequences in GenBank. We will next look at some of the basic kinds of data present in GenBank. Afterward, we will address strategies to extract the data you want from GenBank.

We start with an example. We want to find out the sequence of human retinol-binding protein (RBP4). A fundamental distinction is that both DNA and protein sequences are stored in discrete databases. Furthermore, within each database, sequence data are represented in a variety of forms. For example, RBP4 may be described at the DNA level (as a gene), at the RNA level [as a messenger RNA (mRNA) transcript], and at the protein level (see Fig. 2.2). Because RNA is relatively unstable, it is typically converted to complementary DNA (cDNA), and a variety of databases contain cDNA sequences corresponding to RNA transcripts. Thus for our example of RBP4, the various forms include the following:

RBP4 is also sometimes called RBP. In general, a gene does not always have the same name as the corresponding protein. HD is the gene symbol for the gene that causes Huntingon disease; the gene name corresponds to the phenotype. The gene product (i.e. protein) is called huntingtin. Often, multiple investigators study the same gene or protein and assign different names. The human genome organization (HUGO) Gene Nomenclature Committee (HGNC) has the critical task of assigning official names to genes and proteins. See ▸http://www.gene.ucl.ac.uk/nomenclature/.

Genomic DNA Databases

- RBP4 is part of a chromosome. In the case of human RBP we will see that its gene is situated on chromosome 10 (Chapter 17, on the human genome).
- RBP4 is a part of a large fragment of DNA such as a cosmid, bacterial artificial chromosome (BAC) or yeast artificial chromosome (YAC) that may contain several genes. A BAC is a large segment of DNA [typically up to 200,000 base pairs (bp), or 200 kb] that is cloned into bacteria. Similarly, YACs are used to clone large amount of DNA into yeast. BACs and YACs are useful to sequence large portions of genomes.
- RBP4 is present in databases as a gene. The gene is the functional unit of heredity, and it is a DNA sequence that typically consists of regulatory regions, protein-coding exons, and introns. Often, human genes are 10–50 kb in size.
- RBP4 is present as a sequence-tagged site (STS)—that is, as a small fragment of DNA (typically 500 bp long) that is used to link genetic and physical maps and which is part of a database of sequence-tagged sites (dbSTS).

Chromosome 10, which is a mid-sized chromosome, contains about 1500 genes and is about 150,000 kilobases (kb) in length.

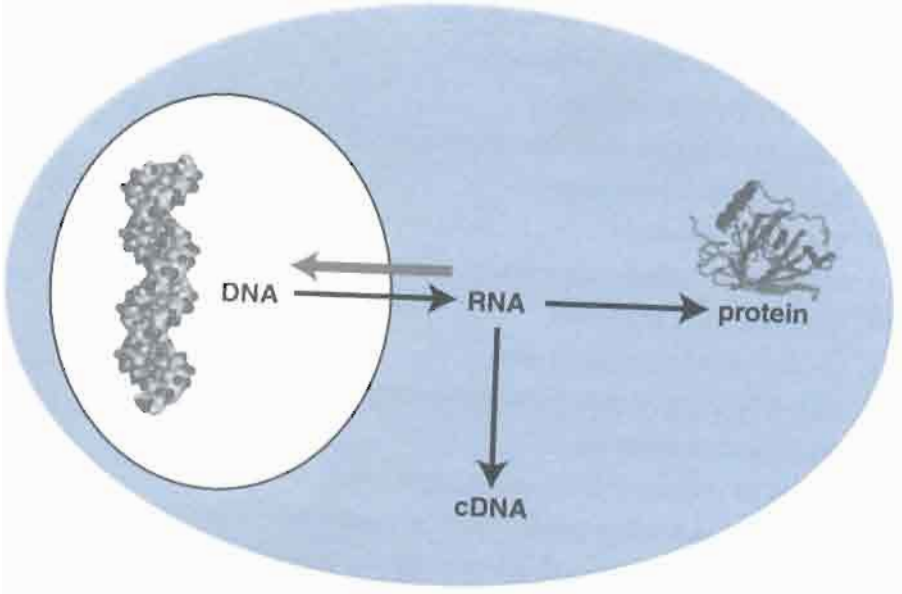


FIGURE 2.2. *Types of sequence data in GenBank and other databases using RBP4 as an example. Note that "retinol-binding protein" may refer to a gene, an RNA transcript (or its corresponding complementary DNA), or a protein. There are specialized databases corresponding to each of these three levels. See text for abbreviations. There are many other databases (not listed) that are not part of GenBank and NCBI; note that SwissProt, PDB, and PIR are protein databases that are independent of GenBank. The raw nucleotide sequence data in GenBank, DDBJ, and EBI are equivalent.*

cDNA Databases Corresponding to Expressed Genes

- RBP4 is represented in databases as an expressed sequence tag (EST), that is, a cDNA sequence derived from a particular cDNA library. If one obtains a tissue such as liver, purifies RNA, then converts the RNA to the more stable form of cDNA, some of the cDNA clones contained in that cDNA are likely to encode RBP4.

Protein Databases

In GenBank, the convention is to use DNA sequences even when referring to RNA.

- The name RBP4 may refer to the DNA, the RNA, or the protein. As a protein, RBP4 is present in databases such as the nonredundant (nr) database of GenBank (Benson et al., 2002), the SwissProt database (Boeckmann et al., 2003), and the Protein Data Bank (Westbrook et al., 2002).

Expressed Sequence Tags (ESTs)

In May, 2003 GenBank had over 16,000,000 ESTs including the "top 10" organisms listed in Table 2.3. Assuming that there are 35,000 human genes (see Chapter 17), there is currently an average of 150 ESTs for each gene.

The database of expressed sequence tags (dbEST) is a division of GenBank that contains sequence data and other information on "single-pass" cDNA sequences from a number of organisms (Boguski et al., 1993). An EST is a partial DNA sequence of a cDNA clone. All cDNA clones, and thus all ESTs, are derived from some specific RNA source such as human brain or rat liver. The RNA is converted into a more stable form, cDNA, which may then be packaged into a cDNA library (refer to Fig. 2.2). ESTs are typically randomly selected cDNA clones that are rapidly sequenced on one strand. ESTs are often 300–800 bp in length. The earliest efforts to sequence ESTs resulted in the identification of many hundreds of genes that were novel at the time (Adams et al., 1991).

Currently, GenBank divides ESTs into three major categories: human, mouse, and other.

TABLE 2-3 Top Ten Organisms for Which ESTs Have Been Sequenced (dbEST release 050903, May 2003)

Many thousands of cDNA libraries have been generated from a variety of organisms, and the total number of public entries is currently over 16 million

Organism	Common Name	Number of ESTs
Homo sapiens	Human	5,142,390
Mus musculus + *domesticus*	Mouse	3,721,428
Rattus sp.	Rat	525,556
Ciona intestinalis	Sea squirt	492,488
Gallus gallus	Chicken	418,093
Triticum aestivum	Wheat	340,945
Hordeum vulgare + subsp.	Barley	340,945
Bos taurus	Cattle	319,775
Danio rerio	Zebrafish	311,335
Glycine max	Soybean	308,582

Source: ▸http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html

ESTs and UniGene

The goal of the UniGene (unique gene) project is to create one unique entry for each gene and to collect all the ESTs associated with that gene. For example, in the case of RBP4, there is only one UniGene entry. The 26 organisms represented in UniGene are listed in Table 2.4.

Although there is only a single UniGene entry for RBP4, this entry currently has over 200 human ESTs that match the RBP gene. This large number of ESTs reflects how abundantly the RBP gene has been expressed in cDNA libraries that have been sequenced. A UniGene cluster is a database entry for a gene containing a group of corresponding ESTs (Fig. 2.3).

UniGene clusters are created by sequence similarity searching (see Chapters 3 and 4) using ESTs, and by gathering aligned sequences into clusters. Many clusters have only one member, representing unique sequences, while other clusters have tens of thousands of EST members. We will discuss UniGene clusters further in Chapter 6 (on gene expression).

There are now thought to be approximately 35,000 human genes (see Chapter 17). One might expect an equal number of UniGene clusters. However, in practice, there are more UniGene clusters than there are genes—currently, there are about 100,000 human UniGene clusters. This discrepancy could occur for three reasons. (1) Clusters of ESTs could correspond to distinct regions of one gene. In that case there would be two (or more) UniGene entries corresponding to a single gene (see Fig. 2.3). As the genome is more fully sequenced, it will become apparent that the two UniGene clusters should properly cluster into one. Thus, the number of UniGene clusters will collapse over time. (2) Some genes may be expressed at very low levels. Currently, 35,000 human UniGene clusters consist of a single EST, and over 60,000 UniGene clusters consist of just one to four ESTs. These could reflect authentic genes that have not yet been appreciated by other means of gene identification. (3) Some DNA may be transcribed during the creation of a cDNA library without corresponding to an authentic transcript. Thus it is a cloning artifact. We will discuss the criteria for defining a eukaryotic gene in Chapters 15–17. Alternative splicing (Chapter 6) may introduce apparently new clusters of genes because the spliced exon has no homology to the rest of the sequence.

To find the entry for RBP, go to ▸http://www.ncbi.nlm.nih.gov, select UniGene on the sidebar, select human, then enter retinol-binding protein or *RBP4*. The UniGene accession number is Hs.76461; note that Hs refers to *H. sapiens*.

We are using *RBP4* as a specific example. If you want to type "retinol-binding protein" as a query, you will simply get more results from any database—in UniGene, you will find several dozen entries.

TABLE 2-4 Organisms Represented in UniGene

Organism	Species	Abbreviation	Common name
Chordata	*Bos taurus*	Bt	Cow
	Ciona intestinalis	Cin	Sea squirt
	Danio rerio	Dr	Zebrafish
	Gallus gallus	Gga	Chicken
	Homo sapiens	Hs	Human
	Mus musculus	Mm	Mouse
	Oryzias latipes	Ola	Japanese medaka
	Rattus norvegicus	Rn	Rat
	Silurana tropicalis	Str	Western clawed frog
	Sus scrofa	Ssc	Pig
	Xenopus laevis	Xl	African clawed frog
Arthropoda	*Anopheles gambiae*	Aga	Malaria mosquito
	Drosophila melanogaster	Dm	Fruit fly
Nematoda	*Caenorhabditis elegans*	Cel	Worm
Embryophyta	*Arabidopsis thaliana*	At	Thale cress
	Glycine max	Gma	Soybean
	Hordeum vulgare	Hv	Barley
	Lycopersicon esculentum	Les	Tomato
	Medicago truncatula	Mtr	Barrel medic
	Oryza sativa	Os	Rice
	Solanum tuberosum	Stu	Potato
	Sorghum bicolor	Sbi	Sorghum
	Triticum aestivum	Ta	Wheat
	Zea mays	Zm	Corn
Chlorophyta	*Chlamydomonas reinhardtii*	Cre	Green alga
Mycetozoa	*Dictyostelium discoideum*	Ddi	Slime mold

Source: ▸ http://www.ncbi.nlm.nih.gov/UniGene/, May 2003.

The dramatic rise in the number of UniGene entries is illustrated in Figure 2.4. Many of the UniGene entries derive from the Cancer Gene Anatomy Project (CGAP), which is described at ▸ http://cgap.nci.nih.gov/.

Sequence-Tagged Sites (STSs)

As of February 2003 there are 226,000 STSs, derived from several commonly studied organisms (Table 2.5).

The dbSTS is an NCBI site containing STSs, which are short genomic landmark sequences for which both DNA sequence data and mapping data are available (Olson et al., 1989). STSs have been obtained from several dozen organisms (Table 2.5). A typical STS is approximately the size of an EST. Because they are sometimes polymorphic, containing short sequence repeats (Chapter 16), STSs can be useful for mapping studies.

Genome Survey Sequences (GSSs)

There are currently 4.6 million GSS entries (almost half human or murine) (February, 2003). This database is accessed via ▸ http://www.ncbi.nlm.nih.gov/dbGSS/index.html.

The GSS division of GenBank is similar to the EST division, except that its sequences are genomic in origin, rather than cDNA (mRNA). The GSS division

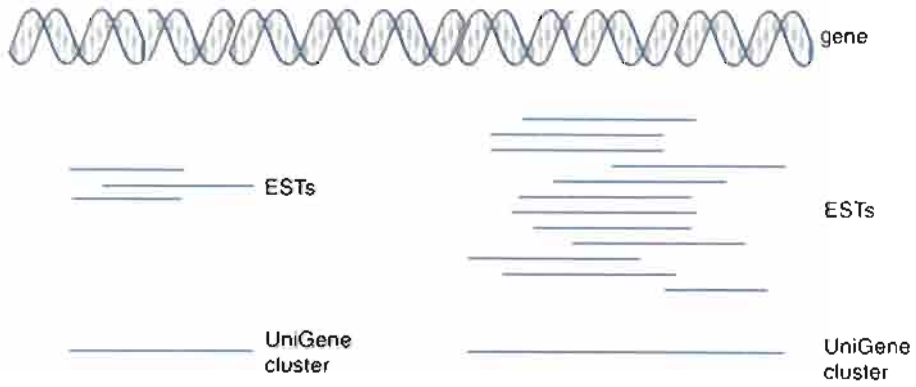


FIGURE 2.3. Schematic description of UniGene clusters. Expressed sequence tags (ESTs) are mapped to a particular gene and to each other. The number of ESTs that constitute a UniGene cluster ranges from 1 to over 1000; on average there are 100 ESTs per cluster. Sometimes, as shown in the diagram, separate UniGene clusters correspond to distinct regions of a gene. Eventually, as genome sequencing increases our ability to define and annotate full-length genes, these two UniGene clusters would be collapsed into one single cluster. Ultimately, the number of UniGene clusters will equal the number of genes in the genome.

contains the following types of data (see Chapters 16 and 17):

- Random "single-pass read" genome survey sequences
- Cosmid/BAC/YAC end sequences
- Exon-trapped genomic sequences
- The *Alu* polymerase chain reaction (PCR) sequences

Recent holdings of the GSS database are listed in Table 2.6.

High-Throughput Genomic Sequence (HTGS)

The HTGS division was created to make "unfinished" genomic sequence data rapidly available to the scientific community. It was done in a coordinated effort between the three international nucleotide sequence databases: DDBJ, EMBL, and GenBank. The HTGS division contains unfinished DNA sequences generated by the high-throughput sequencing centers. The HTGS home page is ▸http://www.ncbi.nlm.nih.gov/HTGS/ and its sequences can be searched via BLAST (see Chapters 4 and 5).

We have described some of the basic kinds of sequence data in GenBank. We will next turn our attention to Entrez and the other programs in NCBI and elsewhere,

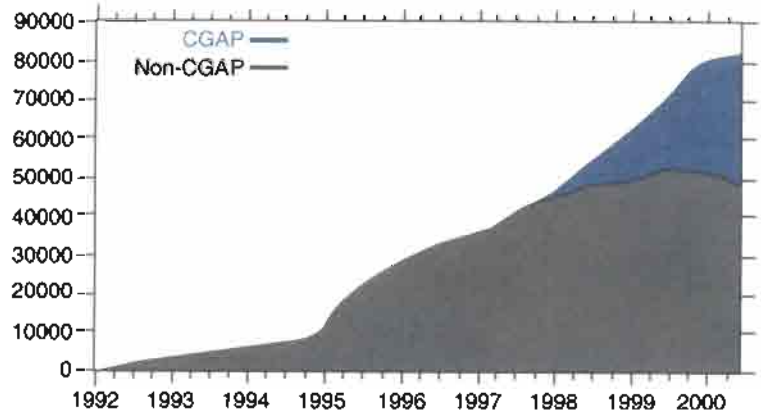


FIGURE 2.4. Growth of human gene entries in UniGene (▸http://www.ncbi.nlm.nih.gov/UniGene/). The Cancer Gene Anatomy Project (CGAP) is described at ▸http://cgap.nci.nih.gov/.

TABLE 2-5 Organisms from Which STSs Have Been Obtained

Organism	Approximate number of STSs
Homo sapiens	134,000
Rattus norvegicus	30,000
Mus musculus	27,000
Danio rerio	22,000
Drosophila melanogaster	1,100

Source: ▸http://www.ncbi.nlm.nih.gov/genome/sts/unists_stats.html,November 2002.

which allow you to access GenBank, EMBL, and DDBJ data and related literature information. In particular, we will introduce the NCBI website, one of the main web-based resources in the field of bioinformatics.

National Center for Biotechnology Information

Introduction to NCBI: Home Page

The NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information (Wheeler et al., 2002). The NCBI home page is shown in Figure 2.5. Across the top bar of the website, there are seven categories: PubMed, Entrez, BLAST, OMIM, Books, Taxonomy, and Structure.

PubMed

PubMed is the search service from the National Library of Medicine (NLM) that provides access to over 11 million citations in MEDLINE (Medical Literature, Analysis, and Retrieval System Online) and other related databases, with links to participating online journals. To access a PubMed tutorial, go to the NCBI home page (▸www.ncbi.nlm.nih.gov), click on "Education" on the left sidebar, then click on the PubMed tutorial button.

Entrez

Entrez integrates the scientific literature, DNA and protein sequence databases, three-dimensional protein structure data, population study data sets, and assemblies of complete genomes into a tightly coupled system. PubMed is the literature component of Entrez.

TABLE 2-6 Selected Organisms from Which GSSs Have Been Obtained

Organism	Approximate Number of STSs
Mus musculus (house mouse)	945,000
Homo sapiens (human)	873,000
Brassica oleracea (vegetable)	567,000
Rattus norvegicus (Norway rat)	307,000
Arabidopsis thaliana (thale cress)	214,000
Tetraodon nigroviridis (pufferfish)	189,000

Source: ▸http://www.ncbi.nlm.nih.gov/dbGSS/dbGSS.summary.html, February 2003.

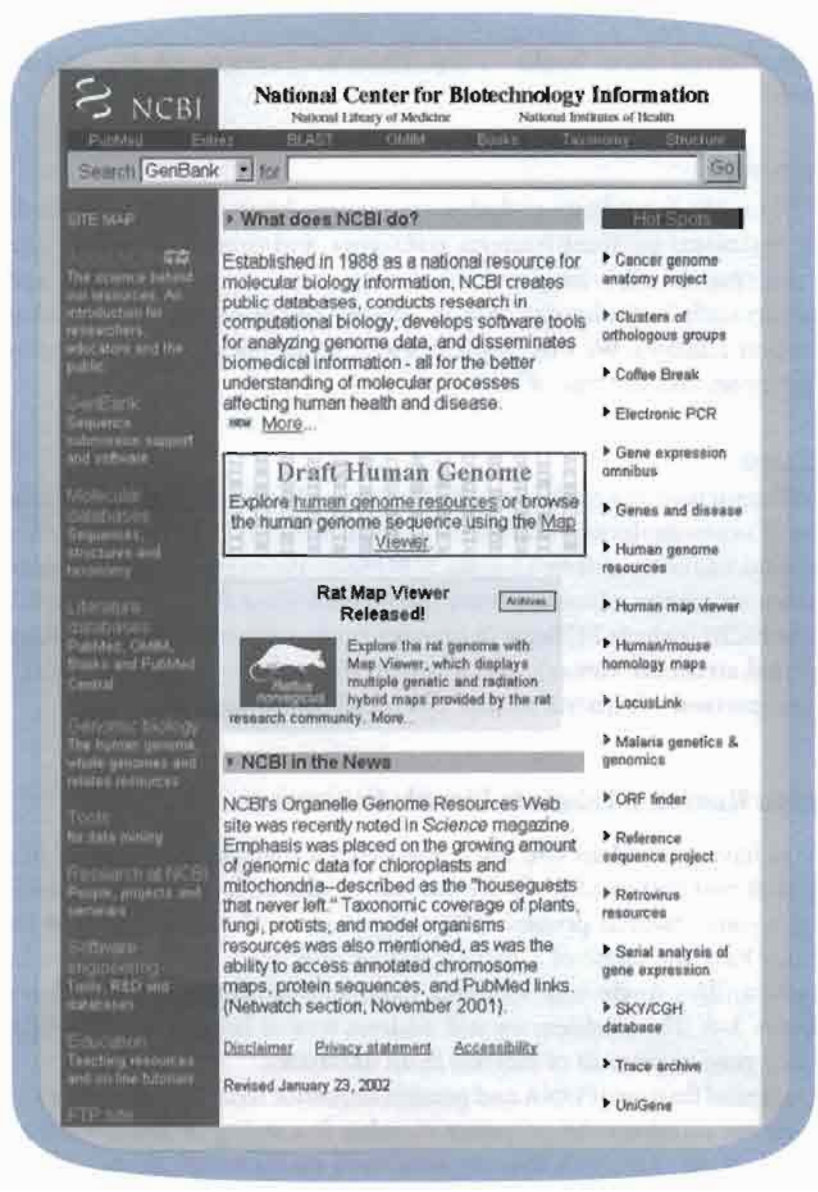


FIGURE 2.5. *The main page of the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov). Across the top bar, sections include PubMed, Entrez and Books (described in this chapter), BLAST (Chapters 3–5), Taxonomy (Chapters 12–17), Structure (Chapter 9), and Online Mendelian Inheritance in Man (OMIM, Chapter 18) Note that the left sidebar includes tutorials within the Education section.*

BLAST

BLAST (Basic Local Alignment Search Tool) is NCBI's sequence similarity search tool designed to support analysis of nucleotide and protein databases (Altschul et al., 1990, 1997). BLAST is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. We will explore BLAST in Chapters 3–5.

OMIM

Online Mendelian Inheritance in Man (OMIM) is a catalog of human genes and genetic disorders. It was created by Victor McKusick and his colleagues and developed for the World Wide Web by NCBI (Hamosh et al., 2002). The database contains detailed reference information. It also contains links to PubMed articles and sequence information. We will describe OMIM in Chapter 18 (on human disease).

Books

NCBI offers several dozen books on-line. These books are searchable, and are linked to PubMed.

Taxonomy

The NCBI taxonomy website includes a taxonomy browser for the major divisions of living organisms (archaea, bacteria, eukaryota, and viruses). The site features taxonomy information such as genetic codes and taxonomy resources and additional information such as molecular data on extinct organisms and recent changes to classification schemes. We will visit this site in Chapters 11 (on evolution) and 12–17 (on genomes and the tree of life).

Structure

The NCBI structure site maintains the Molecular Modelling Database (MMDB), a database of macromolecular three-dimensional structures, as well as tools for their visualization and comparative analysis. MMDB contains experimentally determined biopolymer structures obtained from the Protein Data Bank (PDB). Structure resources at NCBI include PDBeast (a taxonomy site within MMDB), Cn3D (a three-dimensional structure viewer), and a vector alignment search tool (VAST) which allows comparison of structures. (See Chapter 9, on protein structure.)

The Protein Data Bank (http://www.rcsb.org/pdb/) is the single worldwide repository for the processing and distribution of biological macromolecular structure data. We will explore PDB in Chapter 9.

Accession Numbers: Labels to Identify Sequences

When you have a problem you are studying that involves any gene or protein, it is likely that you will need to find information about some database entries. You may begin your research problem with information obtained from the literature or you may have the name of a specific sequence of interest. Perhaps you have raw amino acid and/or nucleotide sequence data; we will explore how to analyze these in Chapters 3–5. The problem we will address now is how to extract information about your gene or protein of interest from databases.

An essential feature of DNA and protein sequence records is that they are tagged with accession numbers. An accession number is a string of about 4–10 numbers and/or alphabetic characters that are associated with a molecular sequence record. An accession number may also label other entries, such as protein structures or even the results of a gene expression experiment (Chapters 6 and 7). Accession numbers from molecules in different databases have characteristic formats (Box 2.1). These formats vary because each database employs its own system. As you explore databases from which you extract DNA and protein data, try to become familiar with the different formats for accession numbers. Some of the various databases (Fig. 2.2) employ accession numbers that tell you whether the entry contains nucleotide or protein data.

DNA is usually sequenced on both strands. However, ESTs are often sequenced on one strand only, and thus they have a high error rate. We will discuss sequencing error rates in Chapter 12.

For a typical molecule such as RBP4 there are many hundreds of accession numbers (Fig. 2.6). Many of these correspond to ESTs and other fragments of DNA that match RBP4. How can you assess the quality of sequence or protein data? Many of the sequence entries contain errors, particularly in the ends of EST reads. When we compare RBP4 sequence derived from mRNA and from genomic DNA, we expect them to match perfectly (or nearly so), but as we will see, there may be discrepancies (e.g., Fig. 6.4). One of the most important recent developments in the management of molecular sequences is RefSeq, which is described below. The goal of RefSeq is to provide the best representative sequence for each normal (i.e., nonmutated)

BOX 2-1
Types of Accession Numbers

Type of Record	Sample Accession Format
GenBank/EMBL/DDBJ nucleotide sequence records	One letter followed by five digits, e.g., X02775 Two letters followed by six digits, e.g., AF025334
GenPept sequence records (which contain the amino acid translations from GenBank/EMBL/DDBJ records that have a coding region feature annotated on them)	Three letters and five digits, e.g., AAA12345
Protein sequence records from SwissProt and PIR	Usually one letter and five digits, e.g., P12345. SwissProt numbers may also be a mixture of numbers and letters.
Protein sequence records from the Protein Research Foundation	A series of digits (often six or seven) followed by a letter, e.g., 1901178A
RefSeq nucleotide sequence records	Two letters, an underscore bar, and six digits, e.g., mRNA records (NM_*): NM_006744; genomic DNA contigs (NT_*): NT_008769
RefSeq protein sequence records	Two letters (NP), an underscore bar, and six digits, e.g., NP_006735
Protein structure records	PDB accessions generally contain one digit followed by three letters, e.g., 1TUP. They may contain other mixtures of numbers and letters (or numbers only). MMDB ID numbers generally contain four digits, e.g., 3973.

Source: Modified from ▸ http://www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html.

A GenBank or RefSeq accession number refers to the most recent version of a given sequence. For example NM_006744.2 is currently a RefSeq identifier for human RBP4. The suffix.2 is the version number. By default, if you do not specify a version number then the most recent version is provided. Try doing an Entrez nucleotide search for NM_006744.1 and you can learn about the revision history of that accession number.

transcript produced by a gene and for each normal protein product (Maglott et al., 2000).

Five Ways to Access DNA and Protein Sequences

How can one navigate through the bewildering number of protein and DNA sequences in the various databases? We will briefly describe five tools: LocusLink (including the important RefSeq project), UniGene, Entrez, Ensembl, and ExPASy. We will use the example of finding RBP4. A common theme is that the various databases are increasingly interconnected, providing a variety of convenient links to each other and to algorithms that are useful for DNA, RNA, and protein analysis.

(1) LocusLink: Centralized Resource for Information on Genes and Proteins

LocusLink is a curated database containing descriptive information about genetic loci (Wheeler et al., 2002; Pruitt et al., 2000). You can obtain information on official nomenclature, aliases, sequence accessions, phenotypes, EC numbers, MIM numbers, UniGene clusters, homology, map locations, and related websites.

LocusLink is accessed from the main NCBI web page or via ▸ http://www.ncbi.nlm.nih.gov/LocusLink/. We will explore many of the resources within LocusLink in later chapters, such as Enzyme Commission (EC) numbers (Chapter 8) and Mendelian Inheritance in Man (MIM, Chapter 18).

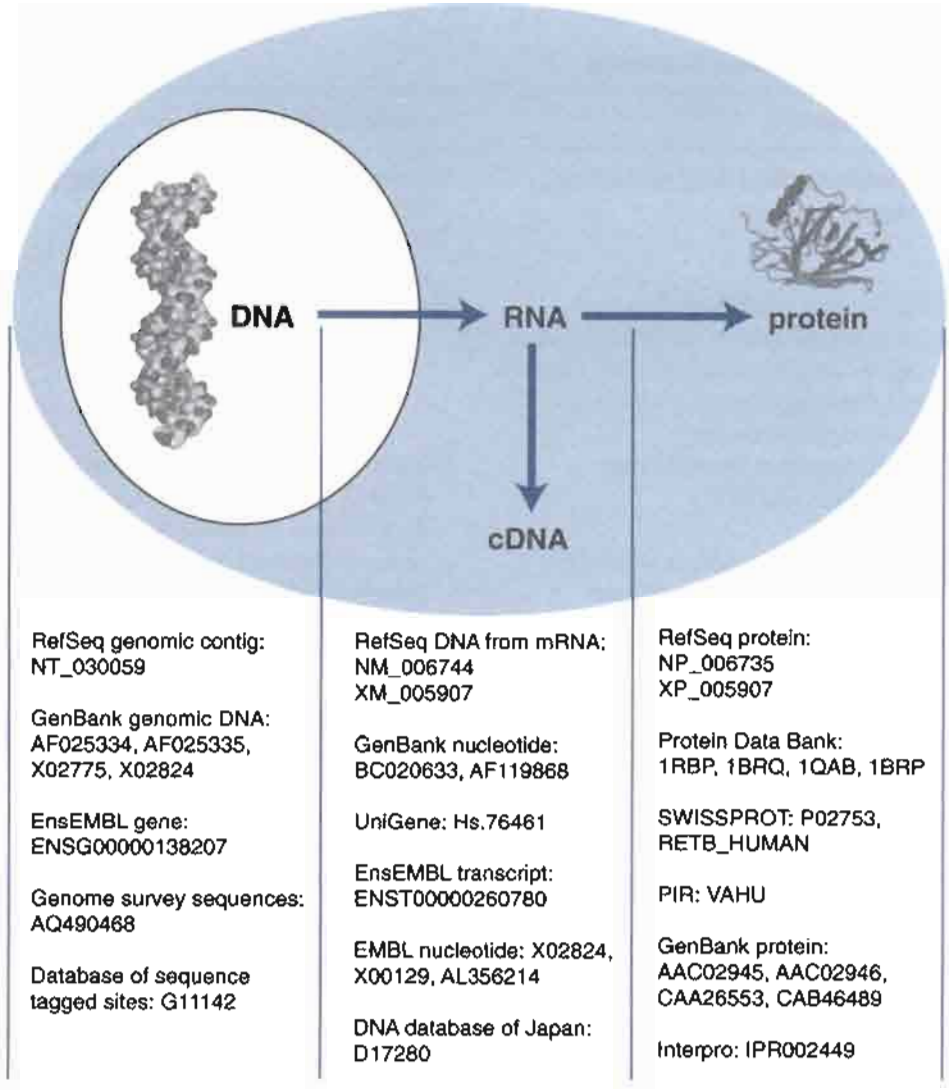


FIGURE 2.6. Examples of accession numbers using RBP4 as an example. Accession numbers are tags that link to the records for sequences derived from genomic DNA (left side of figure), cDNA derived from mRNA (including expressed sequence tag data which are from the sequencing of cDNAs; see middle), and protein data (right side). The RefSeq project is particularly important in trying to provide the best representative sequence of each normal (nonmutated) transcript produced by a gene and of a normal protein sequence. Other accession numbers include the LocusLink LocusID (5950) and OMIM identifier 180250. The database resources listed in this figure are discussed in Chapters 2, 6, and 8.

LocusLink provides a fast, direct way to access the protein or DNA accession number for several species (human, mouse, rat, *Drosophila,* zebrafish, cow, nematode, HIV-1). Of particular interest, LocusLink records display RefSeq entries for each DNA or protein sequence (see below). LocusLink also has links to OMIM, HomoloGene (a database that reports human, mouse, and rat orthologs), assorted mapping information, the Genome Database, and some additional features such as GeneCards (a one-page overview of the main features of a gene).

The result of entering LocusLink and typing "retinol-binding protein" is shown in Figure 2.7. Note that in performing a LocusLink search, it can be convenient to change the default display setting from "brief" to "summary," and it can also be convenient to restrict the search to a particular organism of interest. Clicking on LocusID 5950, the link for the main serum retinol-binding protein RBP4, results in the following (see Fig. 2.8):

- Across the top bar there are direct links to RBP4 entries in various databases such as PubMed, OMIM, UniGene, the NCBI map viewer, and a variation database describing single-nucleotide polymorphisms (SNPs; see Chapter 18).

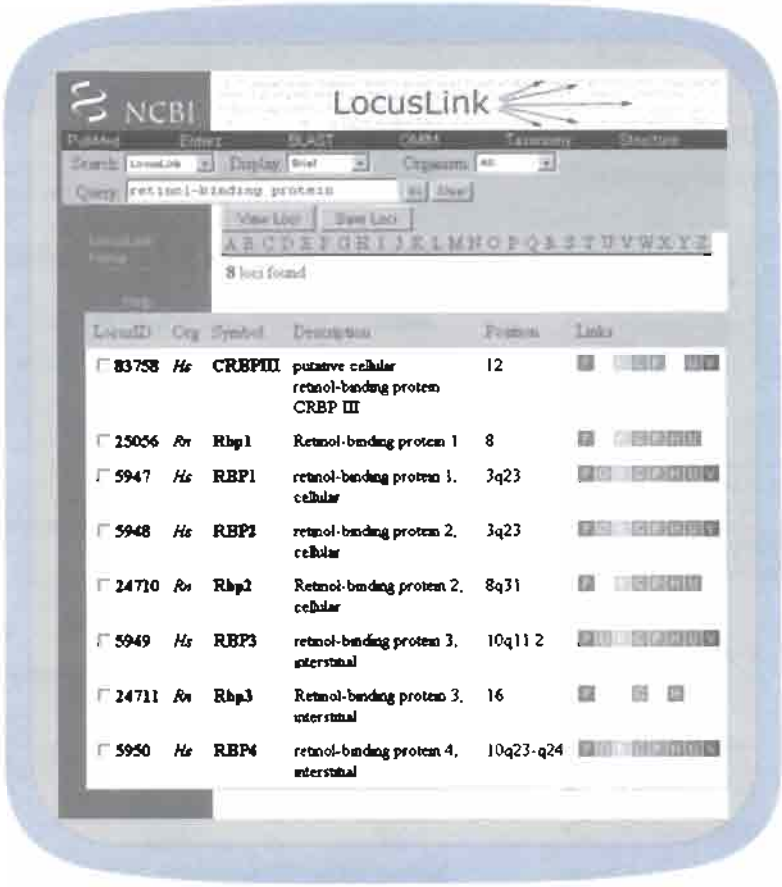


FIGURE 2.7. *Result of a search for "retinol-binding protein" in LocusLink (▸http://www.ncbi.nlm.nih.gov/LocusLink/). Information is provided for organisms, including* H. sapiens *(Hs) and* R. norvegicus *(Rn). Position refers to the chromosomal locus of the gene. A variety of links are provided (see text).*

- There is a brief description of the function of RBP, defining it as a carrier protein of the lipocalin family.
- A section on function further describes the phenotype of RBP deficiency.
- Map information includes the chromosomal locus of the gene (10q23–24).
- The Reference Sequence (RefSeq) accession numbers are provided: NM_006744 for the DNA encoding RBP4 and NP_006735 for the protein entry (see below). GenBank accession numbers corresponding to *RBP4* (both nucleotide and protein) are also provided, including X00129, X02775.

NCBI Reference Sequence (RefSeq): Best Representative Sequences

RefSeq provides one accession number for a gene or protein that corresponds to the most stable, agreed-upon versions of the sequence (Pruitt et al., 2000; Maglott et al., 2000). There may be hundreds of GenBank accession numbers corresponding to a gene, but there will be only one RefSeq entry or several RefSeq entries if there are splice variants. This entry has different status levels (predicted, provisional, and reviewed), but in each case the RefSeq entry is intended to unify the sequence records.

You can recognize a RefSeq accession by its format, such as NP_006735 (for RBP protein) or NM_006744 (for RBP mRNA). In general, RefSeq uses the formats given in Table 2.7.

Allelic variants, such as single base mutations in a gene, are not assigned different RefSeq accession numbers. However, OMIM and dbSNP (Chapter 18) do catalog allelic variants.

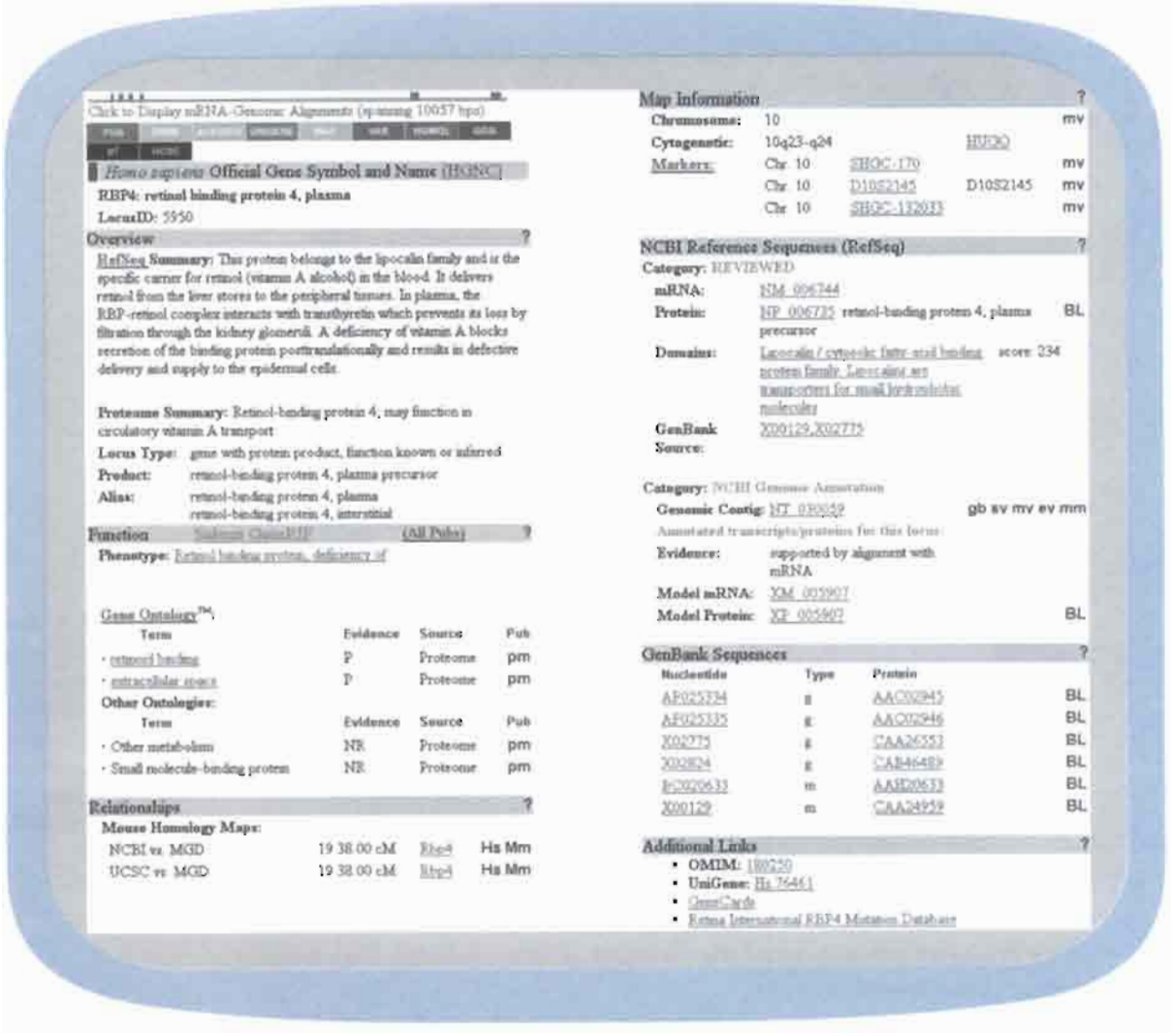


FIGURE 2.8. Portion of the LocusLink entry for human RBP4. *Across the top of the page, there are links to* RBP4 *entries in a variety of databases: PubMed, OMIM, UniGene, Map, a variation database (dbSNP), HomoloGene (with information on mouse and rat* RBP4*), the Genome Database GDB, a gene ontology database, and an Ensembl viewers. We will describe these databases in later chapters. The LocusLink entry provides other key information on* RBP4*, including links to the definitive nucleotide and protein entries.*

TABLE 2-7 Formats of Accession Numbers for RefSeq Entries

Molecule	Accession Format	Genome
Complete genome	NC_######	Archaea, bacterial, organelle, virus
Complete chromosome	NC_######	Eukaryote
Complete sequence	NC_######	Plasmid
Genomic contig	NT_######	*Homo sapiens*
Genomic scaffold	NW_######	Rat
mRNA	NM_######	Limited vertebrate: human, mouse, rat
Protein	NP_######	All of the above

For a description of plasmids see Figure 6.5.

(2) UniGene

As described above, the UniGene project assigns one cluster of sequences to one gene. For example, for *RBP4* there is one UniGene entry with the UniGene accession number Hs.418083. This UniGene entry includes a list of all the GenBank entries, including ESTs, that correspond to the RBP4 gene. The UniGene entry also includes mapping information, homologies, and expression information (i.e., a list of the tissues from which cDNA libraries were generated that contain ESTs corresponding to the RBP gene).

Comparison of UniGene and LocusLink. UniGene and LocusLink have features in common, such as links to OMIM, homologs, and mapping information. They both show RefSeq accession numbers. There are four main differences between UniGene and LocusLink:

1. UniGene has detailed expression information; the regional distributions of cDNA libraries from which particular ESTs have been sequenced are listed.
2. UniGene lists ESTs corresponding to a gene, allowing one to study them in detail.
3. LocusLink may provide a more stable description of a particular gene; as described above, UniGene entries may be collapsed as genome-sequencing efforts proceed.
4. LocusLink has fewer entries than UniGene, but these entries are better curated.

LocusLink now has about 41,000 human RefSeq entries (as of May 2003). Over 14,000 of these are characterized "gene with protein product, function known or inferred."

(3) Entrez

Entrez is a popular tool that is commonly used to extract sequence information. Like SRS (see below) it is a query, retrieval and display system. Try to find human RBP4. From the NCBI main page (▸www.ncbi.nlm.nih.gov) click Entrez or go to Entrez directly (▸http://www.ncbi.nlm.nih.gov/Entrez/). Specify that you want to do a protein search, then type "retinol-binding protein." Over 200 separate results ("hits") are returned, which is too many. Try typing "human retinol-binding protein" and approximately 75 hits are returned. A search for "retinol-binding protein human interstitial"yields nine hits, including one with a RefSeq number (NP_006735). Figure 2.9 shows the standard, default form of this Entrez record. FASTA is another common format for protein (or DNA) sequences, as shown in Figure 2.10. It is simple to obtain a variety of formats by changing the Entrez display options.

Note that your search results are likely to vary as the database changes over time.

FASTA is both an alignment program (described in Chapter 3) and a sequence format (further described in Chapter 4).

Note that extremely useful tutorials are available for Entrez at ▸http://www.ncbi.nlm.nih.gov/Education/index.html. One of the many features of an Entrez search is that limits may be imposed to restrict the output to the information you are seeking. In the case of a search for "RBP," a search of the Entrez nucleotide database returns about 550 entries. By selecting "limits" and changing the pull-down menu to "Only from RefSeq" and then selecting "go," only 65 records are returned. Thus, the information you are seeking is easier to find. In this example, the *H. sapiens RBP4* DNA sequence is easy to find on the list (RefSeq accession NM_006744).

(4) European Bioinformatics Institute and Ensembl

The EBI provides access to sequences via the EMBL nucleotide database (Hingamp et al., 1999). The database can be searched dozens of ways (such as by keyword) through its sequence retrieval system, SRS6 (Etzold et al., 1996). Thus, EBI offers searches comparable to those of the NCBI GenBank database using Entrez. EBI also sponsors Ensembl, a comprehensive website for bioinformatic analysis of the

Ensembl (▸http://www.ensembl.org) is supported by the EBI in cooperation with The Wellcome Trust Sanger Institute (▸http://www.sanger.ac.uk/). EMBL is accessible at ▸http://www.ebi.ac.uk/embl/.

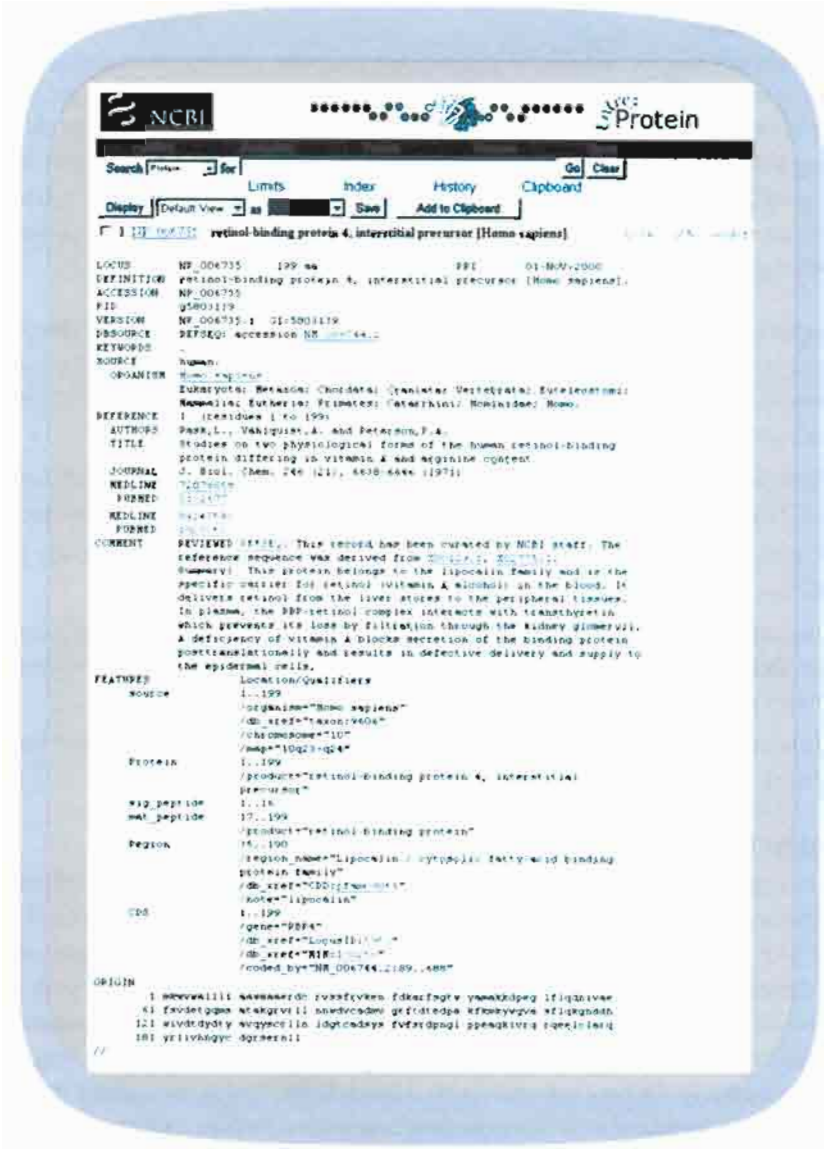


*FIGURE 2.9. Entrez display of a GenBank protein record for human RBP4. This is a typical entry for any protein. (DNA records are similar.) Key information includes the length of the protein (199 amino acids), the division (PRI, or primate), the accession number (NP_006735), the organism (*H. sapiens*), literature references, comments on the function of RBP, and features such as the coding sequence (CDS). The amino acid sequence is provided at the bottom. At the top of the page, the display option allows you to obtain this record in a variety of formats, such as FASTA (Fig. 2.10).*

human genome. A text search for "rbp4" at the Ensembl site yields a link to the RBP4 protein and gene; we will return to the Ensembl resource in later chapters. This entry contains a large number of features relevant to RBP4, including identifiers, the DNA sequence, and convenient links to many other database resources (including OMIM and RefSeq).

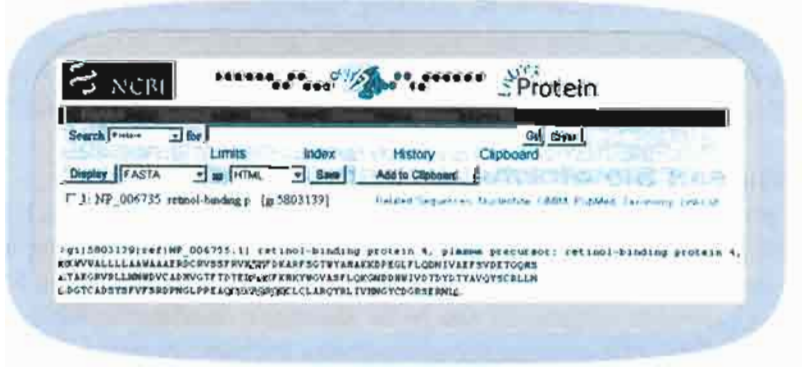


FIGURE 2.10. The protein entry for human RBP4 *can be displayed in the FASTA format. This is easily accomplished by adjusting the "Display" pull-down menu from an Entrez protein record. The FASTA format is used in a variety of software programs that we will use in later chapters.*

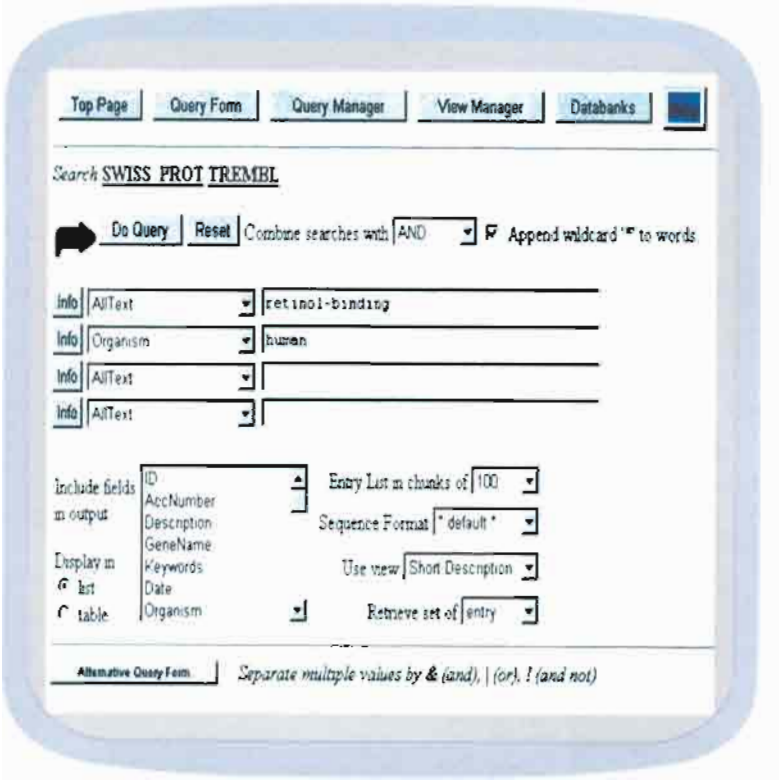


FIGURE 2.11. Format of a query at the Sequence Retrieval System (SRS) of the Expert Protein Analysis System (ExPASy)(▸http://www.expasy.ch/srs5/). This website provides one of the most useful resources for protein analysis. You can also access the SRS through other sites such as the European Bioinformatics Institute (▸http://srs6.ebi.ac.uk/).

(5) ExPASy

One of the most useful resources available to obtain protein sequences and associated data is provided by ExPASy, the Expert Protein Analysis System. The query page has four rectangular boxes (Fig. 2.11). Each has an associated pull-down menu, and as a default condition each says "AllText." In the first box, type "retinol-binding." (Note that queries should consist of one word.) In the second box, type "human," change the corresponding pull-down menu to "organism," then click "do query." You see 10 entries listed. Click the link in which we are interested (SWISS_PROT: RETB_HUMAN P02753).

An output consists of a SwissProt record. This provides very useful, well-organized information, including alternative names and accession numbers; literature links; functional data and information about cellular localization; links to GenBank and other database records for both the RBP protein and gene; and links to many databases such as OMIM, InterPro, Pfam, Prints, GeneCards, PROSITE, and two-dimensional protein gel databases. We will describe these resources later (Chapters 8–10). The record includes features; note that by clicking on any of the linked features, you can see the protein sequence with that feature highlighted in color.

ExPASy is a proteomics server of the Swiss Institute of Bioinformatics (▸http://www.expasy.ch/), another portal from which the Sequence Retrieval System (SRS) is accessed. From ▸http://www.expasy.ch/srs5/, click "Start a new SRS session," then click "continue." SRS was created by Lion Biosciences (▸http://www.lionbioscience.com), and a list of several dozen publicly available SRS servers is at ▸http://downloads.lionbio.co.uk/publicsrs.html.

While we have mentioned several key ways to acquire sequence data, there are dozens of other useful servers. As an example, the Protein Information Resource (PIR) provides access to sequences (Wu et al., 2002). PIR is especially useful for its efforts to annotate functional information on proteins.

Example of How to Access Sequence Data: HIV *pol*

Consider reverse transcriptase, the RNA-dependent DNA polymerase of HIV-1 (Frankel and Young, 1998). The gene-encoding reverse transcriptase is called *pol* (for polymerase). How do you obtain its DNA and protein sequence?

Go to the Entrez protein division of NCBI, set the search to the Genome database, and type "hiv-1." There are two results: HIV-1, and another virus. The accession number for HIV-1 (NC_001802) is a RefSeq number that corresponds to the complete viral genome. Click on the HIV-1 link, and you see a graphic representation

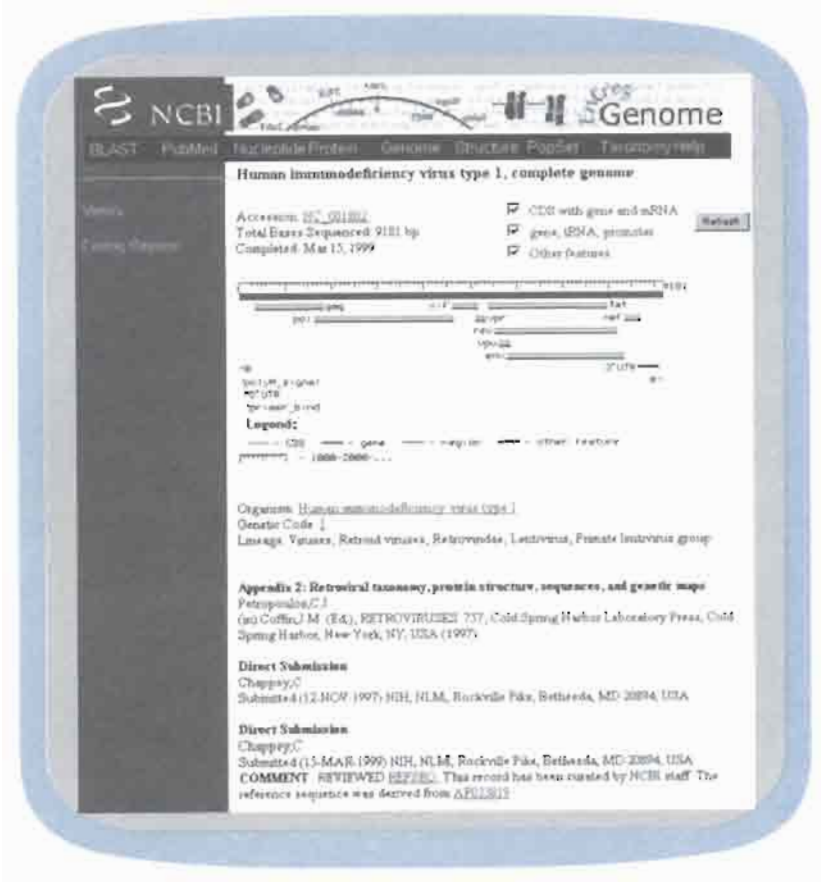


FIGURE 2.12. Entrez genome entry for human immunodeficiency virus 1 (HIV-1) (obtained from ▸ http://www.ncbi.nlm.nih.gov/Entrez/). The RefSeq identifier beginning NC_ refers to a completed genome (Table. 2.6). Key information in this Entrez record includes the accession number, the size of the genome (9181 bp), a graphical overview of the genes/proteins in the genome, taxonomic information, and literature references. On the left sidebar, the "Coding regions" option provides a link to the accession numbers for the nine genes that comprise HIV-1, including pol.

of the nine HIV-1 open reading frames (ORFs) (Fig. 2.12). One of these, *pol,* is the polymerase gene. There are several ways to find its protein sequence. First, click on the NC_001802 link; the protein sequence is available within the Entrez entry, along with the specific protein accession number (NP_057849.2). NP_057849 is the RefSeq identifier, while the suffix ".2" indicates that this is the second version of the sequence that has been entered in GenBank. We may note that this protein has 1003 amino acid residues. As a second approach, use either the "graphical view" or the "coding regions" view on the left sidebar to obtain a graphic view of the proteins in HIV-1. The reverse transcriptase information is readily available this way as well.

In an independent approach to finding this protein sequence, go to the Entrez protein division of NCBI (▸ http://www.ncbi.nlm.nih.gov/Entrez), set the search to the protein database, and type "reverse transcriptase." There are over 40,000 results. To narrow the search results, modify the query to "reverse transcriptase hiv-1." There are now over 15,000 results--still too many. Select "limits," then change the "only from" pull-down menu to "RefSeq" and "Go." There is now a manageable number of search results, including the appropriate NP_057849 accession number.

Note that other NCBI databases are not appropriate for finding the sequence of a viral reverse transcriptase: UniGene does not incorporate viral records, while OMIM is limited to human entries. UniGene and OMIM, however, do have links to genes that are related to HIV, such as eukaryotic reverse transcriptases.

Finally, one can obtain the HIV-1 reverse transcriptase sequence from SRS. Select the SwissProt database to search. In the four available dialog boxes, set one row to "organism" and "HIV-1," then set another row to "AllText" and "reverse." Upon clicking "Do query," a list of several dozen entries is returned; many of these are

PIR is a division of the National Biomedical Research Foundation (▸ http://pir.georgetown.edu/). The February 2003 release contained about 1.1 million entries. PIR was founded by Margaret Dayhoff, whose work is described in Chapter 3.

The United Protein Databases (UniProt) is a new centralized protein resource. It consists of a combined database from PIR, the European Bioinformatics Institute (EBI) and the Swiss Institute of Bioinformatics (SIB). The web site is ▸ http://www.uniprot.org.

You can access the Entrez site at ▸ http://www.ncbi.nlm.nih.gov/Entrez. Note that databases evolve rapidly over time, so your particular search results may vary from those described in the text.

The NLM website is ▸ http://www.nlm.nih.gov/.

identified as fragments and may be ignored. The first entry is SWISS_PROT:POL_HV1A2 (SWISS-PROT accession P03369), a protein of 1003 amino acids. Following the SwissProt link, one finds the "NiceProt" for this database entry. This information includes entry and modification dates, names of this protein and synonyms, references (with PubMed links), comments (including a brief functional description), cross-references to over a dozen other useful databases, a keyword listing, features such as predicted secondary structure, and finally, the amino acid sequence in the single-letter amino acid code and the predicted molecular weight of the protein.

MEDLINE is also accessible through the SRS at the European Bioinformatics Instititute via ▸http://srs.ebi.ac.uk/. A PubMed tutorial is offered at ▸http://www.nlm.nih.gov/bsd/pubmed_tutorial/m1001.html.

Access to Biomedical Literature

The NLM is the world's largest medical library. In 1971 the NLM created MEDLINE (Medical Literature, Analysis, and Retrieval System Online), a bibliographic database. MEDLINE currently contains over 11 million references to journal articles in the life sciences with citations from over 4300 biomedical journals in 70 countries. Free access to MEDLINE is provided on the World Wide Web through PubMed (▸http://www.ncbi.nlm.nih.gov/PubMed/), which is developed by NCBI. While MEDLINE and PubMed both provide bibliographic citations, PubMed also contains links to online full-text journal articles. PubMed also provides access and links to the integrated molecular biology databases maintained by NCBI. These databases contain DNA and protein sequences, genome-mapping data, and three-dimensional protein structures.

The growth of MEDLINE is described at ▸http://www.nlm.nih.gov/bsd/medline_growth.html. Despite the multinational contributions to MEDLINE, the percentage of articles written in English has risen from 59% at its inception in 1966 to 88% in the year 2000 ▸(http://www.nlm.nih.gov/bsd/medline_lang_distr.html).

The National Library of Medicine also offers access to PubMed through NLM Gateway (▸http://gateway.nlm.nih.gov). This comprehensive service includes access to a variety of NLM databases not offered through PubMed, such as meeting abstracts and a medical encyclopedia.

PubMed Central and Movement toward Free Journal Access

The biomedical research community is currently engaged in a debate about the nature of access to literature information. Many journals now offer electronic versions that are available either freely or by paid subscription through the World Wide Web. Groups such as the Association of Research Libraries (ARL) monitor the migration of publications to an electronic form. Thousands of journals are currently available online. Increasingly, online versions of articles include supplementary material such as molecular data (e.g., the sequence of complete genomes, or gene expression data) or videotapes illustrating an article. PubMed Central provides a central repository for biological literature (Roberts, 2001). All these articles have been peer-reviewed and published simultaneously in another journal.

The ARL website is ▸http://www.arl.org/scomm/edir/index.html.

Example of PubMed Search: RBP

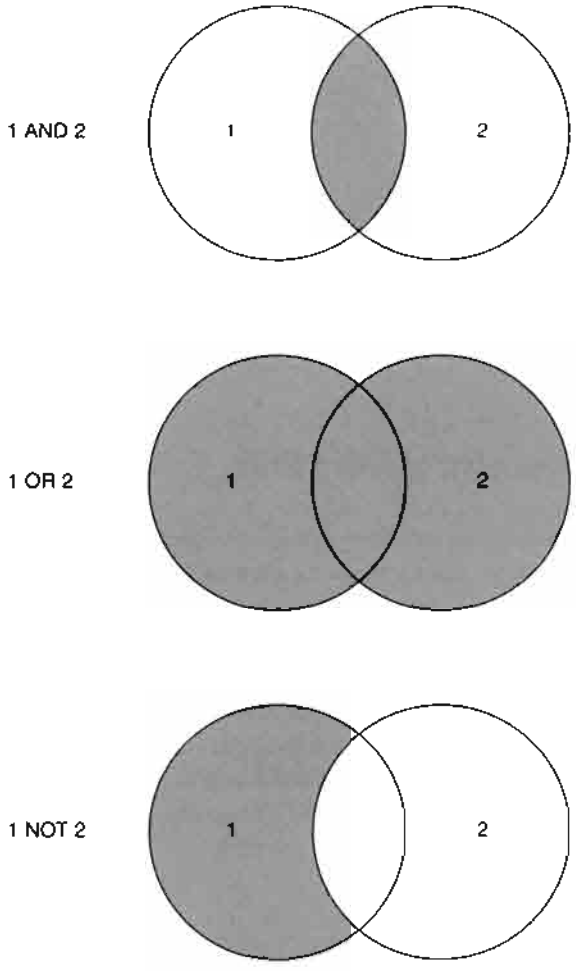
A search of PubMed for information about "RBP" yields 1200 entries. Box 2.2 describes the basics of using Boolean operators in PubMed. There are many additional ways to limit this search. Press "limits" and try applying features such as restricting the output to articles that are freely available through PubMed Central.

The Medical Subject Headings (MeSH) browser provides a convenient way to focus or expand a search. MeSH is a controlled vocabulary thesaurus containing over 22,000 headings. From PubMed, click "MeSH Database" on the left sidebar and enter "retinol-binding protein." The result suggests a series of possibly related topics. By adding MeSH terms, a search can be focused and structured according to the specific information you seek. Lewitter (1998) and Fielding and Powell (2002) discuss strategies for effective MEDLINE searches, such as avoiding inconsistencies in MeSH terminology and finding a balance between sensitivity (i.e., finding relevant articles) and specificity (i.e., excluding irrelevant citations). For example, for a subject that is not well indexed, it is helpful to combine a text keyword with a MeSH

The MeSH website at NLM is ▸http://www.nlm.nih.gov/mesh/meshhome.html.

BOX 2-2
Venn Diagrams of Boolean Operators AND, OR, and NOT for Hypothetical Search Terms 1 and 2

The AND command restricts the search to entries that are both present in a query. The OR command allows either one or both of the terms to be present. The NOT command excludes query results. The gray areas represent search queries that are retrieved. Examples are provided for the queries "lipocalin" or "retinol-binding protein" in PubMed. The Boolean operators affect the searches as indicated.



term. It can also be helpful to use truncations; for example, the search "therap*" introduces a wildcard that will retrieve variations such as therapy, therapist, and therapeutic.

PERSPECTIVE

Bioinformatics is a young, emerging field whose defining feature is the accumulation of biological information in databases. The three major DNA databases—GenBank, EMBL, and DDBJ—are adding several million new sequences each year as well as billions of nucleotides.

In this chapter, we described ways to find information on the DNA and/or protein sequence of RBP4 and the HIV *pol* gene. In addition to the three major databases, a variety of additional resources are available on the web. Increasingly, there is no single correct way to find information—many approaches are possible. Moreover, resources such as those described in this chapter—NCBI, ExPASy, EBI/EMBL, and Ensembl—are closely interrelated, providing links between the databases.

Pitfalls

There are many pitfalls associated with the acquisition of both sequence and literature information. In any search, the most important first step is to define your goal: for example, decide whether you want protein or DNA sequence data. A common difficulty that is encountered in database searches is receiving too much information; this problem can be addressed by learning how to generate specific searches with appropriate limits.

Web Resources

TABLE 2-8 Databases Containing Nucleotide and Protein Information

GenBank	▸http://www.ncbi.nlm.nih.gov/Genbank
DNA Database of Japan (DDBJ)	▸http://www.ddbj.nig.ac.jp/
EMBL/EBI (European Molecular Biology Lab/ European Bioinformatics Institute)	▸http://www.ebi.ac.uk/
Ensembl	▸http://www.ensembl.org/
Protein Information Resource	▸http://pir.georgetown.edu/
SRS at ExPASy	▸http://www.expasy.ch/srs5/
SRS at DDBJ	▸http://srs.ddbj.nig.ac.jp/index-e.html

You can visit the website for this book (▸http://www.bioinfbook.org) to find all the URLs, organized by chapter.

TABLE 2-9 Databases Containing Literature Information

PubMed	▸http://www.ncbi.nlm.nih.gov/PubMed/
Public Library of Science	▸http://www.publiclibraryofscience.org/

Discussion Questions

[2-1] What categories of errors occur in databases? How are these errors assessed?

[2-2] How is quality control maintained in GenBank, given that thousands of individual investigators submit data?

Problems

[2-1] How many human proteins are bigger than 300,000 daltons? (*Hint:* Use Entrez.)

[2-2] You are interested in learning about genes involved in breast cancer. Which genes have been implicated? What are the DNA and protein accession numbers for several of these genes? Try all five of these approaches: PubMed, Entrez, LocusLink, OMIM, and SRS at ExPASy.

[2-3] You would like to know more about *BRCA1*, a tumor suppressor implicated in breast cancer. From which species has this gene been sequenced?

[2-4] An ATP (adenosine triphosphate) binding cassette (ABC) is an example of a common protein domain that is found in many so-called ABC transporter proteins. However, you are not familiar with this motif and would like to learn more.

Approximately how many human proteins have ABC domains? Approximately how many bacterial proteins have ABC domains? Which of the five resources you used in problem 2.2 is most useful in providing you a clear definition of an ABC motif? (We will discuss additional resources to solve this problem in Chapter 8.)

[2-5] Find the accession number of a lipocalin protein (e.g., retinol-binding protein, lactoglobulin, any bacterial lipocalin, glycodelin, or odorant-binding protein). First, use LocusLink, then UniGene, then OMIM, then Entrez nucleotide, and finally Entrez protein. Which approach is most effective? What is the function of this protein?

[2-6] Three main tools for *text*-based searching of molecular information are:

- the National Center for Biotechnology Information's PubMed, Entrez, and OMIM tools (▸ http://www.ncbi.nlm.nih.gov),
- the European Bioinformatics Institute (EBI) Sequence Retrieval System (SRS) (▸ http://srs.ebi.ac.uk) or its related SRS site (▸ http://www.expasy.ch/srs5/), and
- DBGET, the GenomeNet tool of Kyoto University, and the University of Tokyo (▸ http://www.genome.ad.jp/dbget/dbget2.html) literature database LitDB.

You are interested in learning more about West Nile virus. What happens when you use that query to search each of these three resources?

[2-7] You would now like to know what articles about viruses have been published in the journal *Bioinformatics*. Do this search using PubMed.

SELF-TEST QUIZ

[2-1] Which of the following is a RefSeq accession number corresponding to an mRNA?
(a) J01536
(b) NM_15392
(c) NP_52280
(d) AAB134506

[2-2] Approximately how many human clusters are currently in UniGene?
(a) About 8,000
(b) About 40,000
(c) About 100,000
(d) About 150,000

[2-3] You have a favorite gene, and you want to determine in what tissues it is expressed. Which one of the following resources is likely the most direct route to this information?
(a) UniGene
(b) Entrez
(c) LocusLink
(d) PCR

[2-4] Which of the following organisms is NOT represented in LocusLink?
(a) Mouse
(b) Fly
(c) Human
(d) *Escherichia coli*

[2-5] Is it possible for a single gene to have more than one UniGene cluster?
(a) Yes
(b) No

[2-6] Which of the following databases is derived from mRNA information?
(a) dbEST
(b) PBD
(c) OMIM
(d) HTGS

[2-7] Which of the following databases can be used to access text information about human diseases?
(a) EST
(b) PBD
(c) OMIM
(d) HTGS

[2-8] What is the difference between RefSeq and GenBank?
(a) RefSeq includes publicly available DNA sequences submitted from individual laboratories and sequencing projects.
(b) GenBank provides nonredundant curated data.
(c) GenBank sequences are derived from RefSeq.
(d) RefSeq sequences are derived from GenBank and provide nonredundant curated data.

[2-9] If you want literature information, what is the best website to visit?
(a) OMIM
(b) Entrez
(c) PubMed
(d) PROSITE

[2-10] Compare the use of Entrez and ExPASy to retrieve information about a protein sequence.
(a) Entrez is likely to yield a more comprehensive search because GenBank has more data than EMBL.
(b) The search results are likely to be identical because the underlying raw data from GenBank and EMBL are the same.
(c) The search results are likely to be comparable, but the SwissProt record from ExPASy will offer a different output format with distinct kinds of information.

Suggested Reading

Bioinformatics databases are evolving extremely rapidly. Each January, the first issue of the journal *Nucleic Acids Research* includes nearly 100 brief articles on databases. These include descriptions of NCBI (Wheeler et al., 2002), GenBank (Benson et al., 2002), and EMBL (Stoesser et al., 2002). Of particular interest is a description of the Molecular Biology Database Collection by Andreas Baxevanis (2002). This collection is notable for its extraordinary breadth. As molecular sequence data relating to DNA, RNA, and protein are acquired, many specialized databases are created to analyze the structure and function of molecules, cells, tissues, and organisms. The Molecular Biology Database Collection currently lists almost 300 databases (▸http://nar.oupjournals.org/cgi/content/full/31/1/1/DC1).

References

Adams, M. D., et al. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252,** 1651–1656 (1991).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

Altschul, S. F., et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).

Baxevanis, A. D. The Molecular Biology Database Collection: 2002 update. *Nucleic Acids Res.* **30,** 1–12 (2002).

Benson, D. A., et al. GenBank. *Nucleic Acids Res.* **30,** 17–20 (2002).

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31,** 365–370 (2003).

Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. dbEST—database for "expressed sequence tags." *Nat. Genet.* **4,** 332–333 (1993).

Etzold, T., Ulyanov, A., and Argos, P. SRS: Information retrieval system for molecular biology data banks. *Methods Enzymol.* **266,** 114–128 (1996).

Fielding, A. M., and Powell, A. Using Medline to achieve an evidence-based approach to diagnostic clinical biochemistry. *Ann. Clin. Biochem.* **39,** 345–350 (2002).

Fleischmann, R. D., et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).

Frankel, A. D., and Young, J. A. HIV-1: Fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67,** 1–25 (1998).

Hamosh, A., et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **30,** 52–55 (2002).

Hingamp, P., van den Broek, A. E., Stoesser, G., and Baker, W. The EMBL Nucleotide Sequence Database. Contributing and accessing data. *Mol. Biotechnol.* **12,** 255–267 (1999).

Lewitter, F. Text-based database searching. *Bioinformatics: A Trends Guide* **1998,** 3–5 (1998).

Maglott, D. R., Katz, K. S., Sicotte, H., and Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res.*, 28, 126–128 (2000).

O'Donovan, C., et al. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief. Bioinform.* **3,** 275–284 (2002).

Olson, M., Hood, L., Cantor, C., and Botstein, D. A common language for physical mapping of the human genome. *Science* **245,** 1434–1435 (1989).

Pruitt, K. D., Katz, K. S., Sicotte, H., and Maglott, D. R. Introducing RefSeq and LocusLink: Curated human genome resources at the NCBI. *Trends Genet.* **16,** 44–47 (2000).

Roberts, R. J. PubMed Central: The GenBank of the published literature. *Proc. Natl. Acad. Sci. USA* **98,** 381–382 (2001).

Sarton, G. *Appreciation of ancient and medieval science during the renaissance (1450–1600).* A. S. Barnes & Co., New York, 1955.

Singer, C. *The Fasciculo di Medicina Venice 1493.* R. Lier and Co., Florence, 1925.

Stoesser, G., et al. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **30,** 21–26 (2002).

Westbrook, J., et al. The Protein Data Bank: Unifying the archive. *Nucleic Acids Res.* **30,** 245–248 (2002).

Wheeler, D. L., et al. Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.* **30,** 13–16 (2002).

Wu, C. H., et al. The Protein Information Resource: An integrated public resource of functional annotation of proteins. *Nucleic Acids Res.* **30,** 35–37 (2002).

Adrenocorticotropin (ACTH)

The complete amino acid sequences are known for corticotropins isolated from the anterior pituitary glands of three different species, pig, beef, and sheep. The structure of sheep ACTH was discussed in the last chapter, and the sequences shown in Table 9 include only those areas of the three molecules where differences are to be found. Although some difference between the content of amide nitrogen groups has been reported for the three species, these are not included in the figure since it has not been possible to rule out, with certainty, the possibility that these variations are due, in part, to the rigors of the isolation and purification techniques employed.

TABLE 9
Variations in Amino Acid Sequences Among Different Preparations of ACTH

		Residue No.								
Preparation	Species	25	26	27	28	29	30	31	32	33
β-Corticotropin	sheep, beef*	Ala.	Gly.	Glu.	Asp.	Asp.	Glu	Ala.	Ser.	Glu.NH_2
Corticotropin A	pig	Asp.	Gly.	Ala.	Glu.	Asp.	Glu	Leu.	Ala.	Glu

* Identity with sheep hormone not absolutely certain but very probable as judged from the nearly complete sequence analysis by J. S. Dixon and C. H. Li (personal communication to the author).

Two points are of particular interest in regard to the sequences shown. First, the corticotropins of sheep and beef are identical and differ from that of the pig. This finding is consonant with the closer phylogenetic relationship of sheep and cows to each other than of either to pigs. Second, chemical differences are found only in that portion of the ACTH molecule which has been shown to be unessential for hormonal activity. Genetic mutations leading to such differences might, therefore, not be expected to impose significant disadvantages in terms of survival, and these genes could become established in the gene pools of the species.

Melanotropin (MSH)

Melanotropin, like the other hormones considered in this chapter, is a typically chordate polypeptide. Indeed, the demonstration of melanocyte-stimulating activity in extracts of tunicates constitutes an

Pairwise alignment involves matching two protein or DNA sequences. The first proteins that were sequenced include insulin (by Frederick Sanger and colleagues; see Figure 11.1) and globins. This figure is from The Molecular Basis of Evolution *by the Nobel laureate Christian Anfinsen (1959, p. 153). It shows the results of a pairwise alignment of a portion of adrenocorticotropic hormone (ACTH) from sheep or cow (top) with that of pig (below). Such alignments, performed manually, led to the realization that amino acid sequences of proteins reflect the phylogenetic relatedness of different species. Furthermore, pairwise alignments reveal the portions of a protein that may be important for its biological function. Used with permission.*

3

Pairwise Sequence Alignment

INTRODUCTION

One of the most basic questions about a gene or protein is whether it is related to any other gene or protein. Relatedness of two proteins at the sequence level suggests that they are homologous. Relatedness also suggests that they may have common functions. By analyzing many DNA and protein sequences, it is possible to identify domains or motifs that are shared among a group of molecules. These analyses of the relatedness of proteins and genes are accomplished by aligning sequences. As we complete the sequencing of many organisms' genomes, the task of finding out how proteins are related within an organism and between organisms becomes increasingly fundamental to our understanding of life.

Two genes (or proteins) are homologous if they have evolved from a common ancestor.

In this chapter we will introduce pairwise sequence alignment. We will adopt an evolutionary perspective in our description of how amino acids (or nucleotides) in two sequences can be aligned and compared. We will then describe algorithms and programs for pairwise alignment.

Protein Alignment: Often More Informative Than DNA Alignment

Given the choice of aligning a DNA sequence or the sequence of the protein it encodes, it is usually more informative to compare protein sequence. There are several reasons for this. Many changes in a DNA sequence (particularly at the third position of a codon) do not change the amino acid that is specified. Furthermore, many amino acids share related biophysical properties (e.g., lysine and arginine are both

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

basic amino acids). The important relationships between related (but mismatched) amino acids in an alignment can be accounted for using scoring systems (described in this chapter). DNA sequences are less informative in this regard. Protein sequence comparisons can identify homologous sequences from organisms that last shared a common ancestor over 1 billion years ago (BYA) (e.g., glutathione transferases) (Pearson, 1996). In contrast, DNA sequence comparisons typically allow lookback times of up to about 600 million years ago (MYA).

When a nucleotide coding sequence is analyzed, it is often preferable to study its translated protein. In Chapter 4 (on BLAST searching), we will see that we can move easily between the worlds of DNA and protein. For example, the tblastn tool from the NCBI BLAST website allows one to search with a protein sequence for related proteins derived from a DNA database (see Chapter 4). This query option is accomplished by translating each DNA sequence into all of the six proteins that it potentially encodes.

Nevertheless, in many cases it is appropriate to compare nucleotide sequences. This comparison can be important in confirming the identity of a DNA sequence in a database search, in searching for polymorphisms, in analyzing the identity of a cloned cDNA fragment, or in many other applications.

Definitions: Homology, Similarity, Identity

Let us consider the lipocalin family of proteins. We will begin with human RBP4 (accession number NP_006735) and bovine β-lactoglobulin (accession number P02754) as two proteins that are distantly but significantly related. The accession number of *RBP4* is obtained from LocusLink, while β-lactoglobulin is obtained from Entrez because this sequence is not currently available in LocusLink.

Two sequences are *homologous* if they share a common evolutionary ancestry. There are no degrees of homology; sequences are either homologous or not (Reeck et al., 1987; Tautz, 1998). Homologous proteins almost always share a significantly related three-dimensional structure. RBP and β-lactoglobulin have very similar structures as determined by X-ray crystallography (Fig. 3.1). When two sequences are homologous, their amino acid or nucleotide sequences usually share significant identity. Thus, while homology is an inference (sequences are homologous or not),

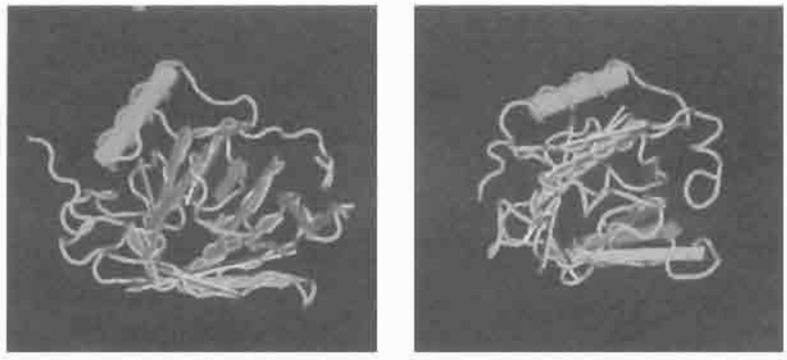


FIGURE 3.1. *Three-dimensional structures of two lipocalins: bovine RBP (left panel), bovine β-lactoglobulin (right panel). The images were generated with the program Cn3D (see Chapter 9). These proteins are homologous (descended from a common ancestor), and they share very similar three-dimensional structures consisting of a binding pocket for a ligand and eight antiparallel beta sheets. However, pairwise alignment of these proteins' amino acid sequences reveals that the proteins share very limited amino acid identity. The accession numbers are MMDB Id: 934 PDB Id: 1FEM (RBP), MMDB Id: 11969 PDB Id: 1BSQ (β-lactoglobulin).*

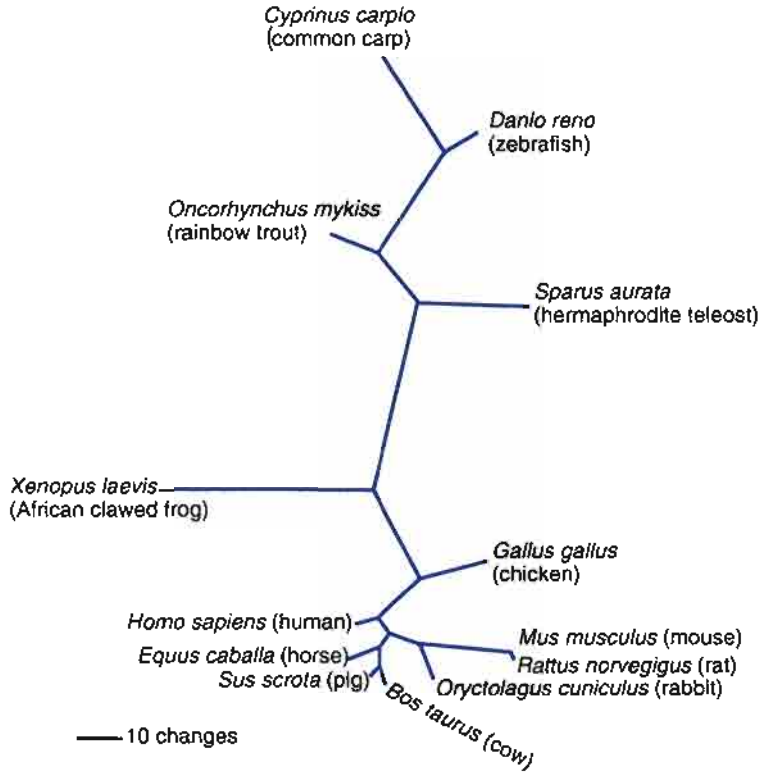


FIGURE 3.2. *Orthologous RBPs. This unrooted tree was generated from a multiple sequence alignment of 13 lipocalins using the PAUP software package (see Chapter 11). The accession numbers of the proteins are as follows:* Homo sapiens *(NP.006735);* Equus caballa *(146257);* Sus scrofa *(A39486);* Bos taurus *(P18902);* Oryctolagus cuniculus *(P06912);* Mus musculus *(Q00724);* Rattus norvegicus *(P04916);* Gallus gallus *(P41263);* Oncorhynchus mykiss *(P24774);* Xenopus laevis *(P06172);* Danio rerio *(CAB64947);* Cyprinus carpio *(CAC12738);* Sparus aurata *(AAF79021). In this tree, sequences that are more closely related to each other are grouped closer together. Note that as entire genomes continue to be sequenced (Chapters 12–17), the number of known orthologs will grow rapidly for most families of orthologous proteins.*

identity and similarity are quantities that describe the relatedness of sequences. Notably, two molecules may be homologous without sharing statistically significant amino acid (or nucleotide) identity. For example, in the lipocalin family, all the members are homologous, but some have sequences that have diverged so greatly that they share no recognizable sequence identity. In general, three-dimensional structures diverge much more slowly than amino acid sequence identity between two proteins (Chothia and Lesk, 1986). Recognizing this type of homology is an especially challenging bioinformatics problem.

Proteins that are homologous may be orthologous or paralogous. *Orthologs* are homologous sequences in different species that arose from a common ancestral gene during speciation. Figure 3.2 shows a tree of RBP orthologs. There is a human RBP gene and a rat gene. Humans and rodents diverged about 75 MYA, at which time a single ancestral RBP gene diverged by speciation. Orthologs are presumed to have similar biological functions; in this example, human and rat RBPs both transport vitamin A in serum. *Paralogs* are homologous sequences that arose by a mechanism such as gene duplication. For example, human plasma RBP is homologous to another carrier protein, human apolipoprotein D (NP.001638) (Fig. 3.3). These two proteins are paralogs. All of the lipocalins have distinct distributions in the body and are thought to have distinct but related functions as carrier proteins.

Walter M. Fitch (1970, p. 113) defined these terms. He wrote that "there should be two subclasses of homology. Where the homology is the result of gene duplication so that both copies have descended side by side during the history of an organism (for example, α and β hemoglobin) the genes should be called paralogous (para = in parallel). Where the homology is the result of speciation so that the history of the gene reflects the history of the species (for example α hemoglobin in man and mouse) the genes should be called orthologous (ortho = exact)."

We thus define homologous genes within the same organism as paralogous. But consider the case of globins. Human α-globin and β-globin are paralogs, as are mouse α-globin and mouse β-globin. Human α-globin and mouse α-globin are orthologs. What is the relation of human α-globin to mouse β-globin? These could be considered paralogs, because α-globin and β-globin originate from a gene duplication event rather than from a speciation event. However, they are not paralogs because they do not occur in the same species. It may thus be most appropriate to simply call them "homologs," reflecting their descent from a common ancestor. Fitch (1970, p. 113) notes that phylogenies require the study of orthologs (see also Chapter 11).

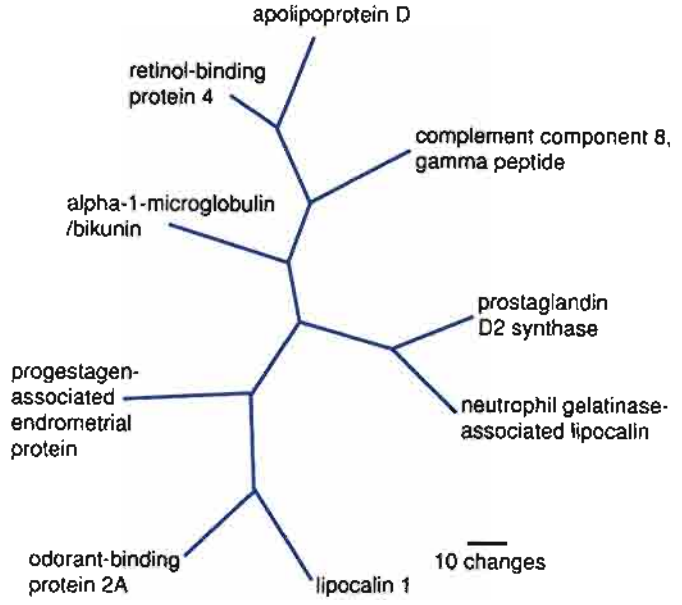


FIGURE 3.3. *Paralogous human lipocalins: Each of these proteins is human, and each is a member of the lipocalin family. This unrooted tree was generated using PAUP (see Chapter 11). The accession numbers of the proteins are: α-1-microglobulin/bikunin (NP_001624); apolipoprotein D (NP_001638); complement component 8, gamma (NP_000597); lipocalin 1 (NP_002288); neutrophil gelatinase-associated lipocalin (JC2339); odorant-binding protein 2A (NP_055397); progestagen-associated endometrial protein (XP_005360); prostaglandin D2 synthase (NP_000945); RBP4 (NP_006735).*

Richard Owen (1804–1892) was one of the first biologists to use the term homology. He defined homology as "the same organ in different animals under every variety of form and function" (Owen, 1843, p. 379).

Notably, orthologs do not necessarily have the same function. We will provide various definitions of gene and protein function in Chapter 8. Later we will explore genomes across the tree of life (Chapters 12–17). In all genome sequencing projects, orthologs are identified based on database searches. Two DNA (or protein) sequences are defined as homologous based on achieving significant alignment scores, as will be discussed in Chapter 4. However, homologous proteins do not necessarily share the same function.

We can assess the relatedness of any two proteins by performing a *pairwise alignment*. In this procedure, we place the two sequences directly next to each other. The sequences of RBP and β-lactoglobulin are shown in Figure 3.4. Note that we use the single-letter amino acid code. The abbreviations are shown in Box 3.1.

It is extremely difficult to align these two proteins by visual inspection. Also, if we allow gaps in the alignment to account for deletions or insertions in the two sequences, the number of possible alignments rises exponentially. Clearly, we will need a computer algorithm to perform an alignment (see Box 3.2). A pairwise alignment of the two proteins is illustrated in Figure 3.5. In this alignment, RBP is on top and β-lactoglobulin is below. A line (|) indicates the presence of an *identical* amino acid in the alignment. For example, notice that along the top row the residues GTWY are all identical between the two proteins. We can count the number of identical residues; in this case, the two proteins share 26% identity (43 residues). Identity is the extent to which two amino acid (or nucleotide) sequences are invariant.

The alignment in Figure 3.5 is constructed using the GAP program of the Genetics Computer Group (GCG). We will describe several freely available web-based tools for pairwise alignment as well as commercial tools such as GCG, later in this chapter.

Later in this chapter we will see that similar pairs of residues have a positive value in a substitution matrix that is used to score an alignment.

Homo sapiens RBP (199 amino acids; accession NP_006735)

```
1    MKWVWALLLL AAWAAAERDC RVSSFRVKEN FDKARFSGTW YAMAKKDPEG
51   LFLQDNIVAE FSVDETGQMS ATAKGRVRLL NNWDVCADMV GTFTDTEDPA
101  KFKMKYWGVA SFLQKGNDDH WIVDTDYDTY AVQYSCRLLN LDGTCADSYS
151  FVFSRDPNGL PPEAQKIVRQ RQEELCLARQ YRLIVHNGYC DGRSERNLL
```

Bos taurus β-lactoglobulin (178 amino acids; accession P02754)

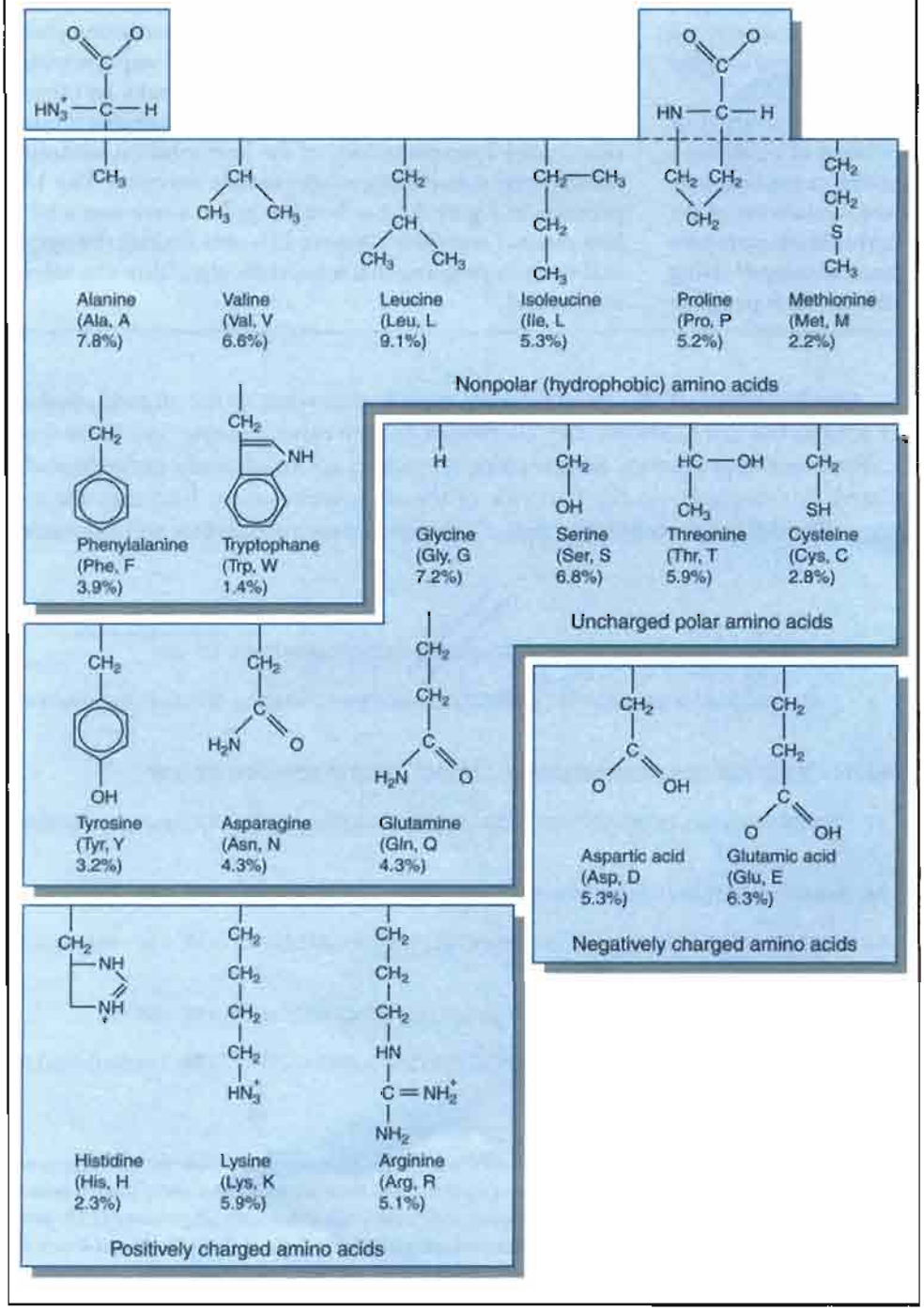
```
1    MKCLLLALAL TCGAQALIVT QTMKGLDIQK VAGTWYSLAM AASDISLLDA
51   QSAPLRVYVE ELKPTPEGDL EILLQKWENG ECAQKKIIAE KTKIPAVFKI
101  DALNENKVLV LDTDYKKYLL FCMENSAEPE QSLACQCLVR TPEVDDEALE
151  KFDKALKALP MHIRLSFNPT QLEEQCHI
```

FIGURE 3.4. *Sequences of RBP and β-lactoglobulin using the standard one-letter amino acid code.*

BOX 3-1

Structures and One- and Three-Letter Abbreviations of Twenty Common Amino Acids

It is very helpful to memorize these abbreviations and to become familiar with the physical properties of the amino acids. The percentages refer to the relative abundance of each amino acid in proteins.



BOX 3-2
Algorithms and Programs

An *algorithm* is a procedure that is structured in a computer program (Sedgewick, 1988). For example, there are many algorithms used for pairwise alignment. A computer *program* is a set of instructions that uses an algorithm (or multiple algorithms) to solve a task. For example, the BLAST program (Chapters 3-5) uses a set of algorithms to perform sequence alignments. Other programs that we will introduce in Chapter 11 use algorithms to generate phylogenetic trees.

Computer programs are essential to solve a variety of bioinformatics problems because millions of operations may need to be performed. The algorithm used by a program provides the means by which the operations of the program are automated. Throughout this book, note how many hundreds of programs have been developed using many hundreds of different of algorithms. Each program and algorithm is designed to solve a specific task. An algorithm that is useful to compare one protein sequence to another may not work in a comparison of one sequence to a database of 10 million protein sequences.

Why is that an algorithm that is useful for comparing two sequences cannot be used to compare millions of sequences? Some problems are so inherently complex that an exhaustive analysis would require a computer with enormous memory or the problem would take an unacceptably long time to complete. A *heuristic algorithm* is one that makes approximations of the best solution without exhaustively considering every possible outcome. The 13 proteins in Figure 3.2 can be arranged in a tree over a billion distinct ways (see Chapter 11)—and finding the optimal tree is a problem that a heuristic algorithm can solve in a second.

Another aspect of this pairwise alignment is that some of the aligned residues are similar but not identical; they are related to each other because they share similar biochemical properties. *Similar* pairs of residues are structurally or functionally related. For example, on the first row of the alignment we can find arginine and lysine (R and K connected by two dots, :); also we can see an aspartate and a glutamate

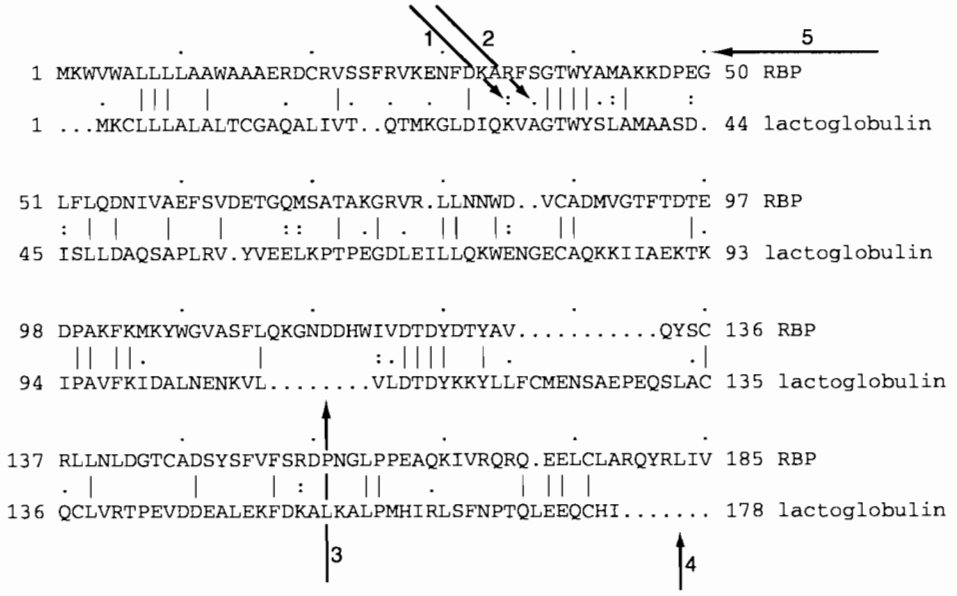


FIGURE 3.5. *Pairwise alignment of human RBP and bovine β-lactoglobulin. Note that the alignment is global (i.e., the entire lengths of each protein are compared), and there are many positions of identity between the two sequences (indicated with bars, |). There are five different kinds of dots in this alignment. (1) The paired dots between the aligned residues indicate different amounts of similarity (e.g., on the top line R and K have two dots and share similar physiochemical properties) (see arrow 1). (2) Single dots between the aligned residues (e.g., arrow 2) also indicate similarity, but less than for paired dots. (3, 4) The alignment contains both internal gaps (indicated by dots in place of alphabetic characters along the sequence; e.g., arrow 3) and terminal gaps at the amino and carboxy termini of β-lactoglobulin (e.g., arrow 4). (5) A dot is indicated above the sequences to mark every 10 bp (arrow 5).*

residue that are aligned. These are *conservative substitutions.* Amino acids with similar properties include the basic amino acids (K, R, H), acidic amino acids (D, E), hydroxylated amino acids (S, T), and hydrophobic amino acids (W, F, Y, L, I, V, M, A).

The *percent similarity* of two protein sequences is the sum of both identical and similar matches. On the top part of Figure 3.5, there are 44 aligned amino acid residues of which 11 are identical and 3 are similar. The percent identity is 25% (11/44) and the percent similarity is 32% (14/44). In general, it is more useful to consider the identity shared by two protein sequences, rather than the similarity, because the similarity measure may be based upon a variety of definitions of how related (similar) two amino acid residues are to each other. We will see in Chapter 4 that database searches using BLAST provide pairwise alignments with both percent identities and "positives" (percent similarities) (Fig. 4.6).

In summary, pairwise alignment is the process of lining up two sequences to achieve maximal levels of identity (and maximal levels of conservation in the case of amino acid alignments). The purpose of a pairwise alignment is to assess the degree of similarity and the possibility of homology between two molecules. We may say that two proteins share 26% amino acid identity or (as in the alignment above) that they share 32% similarity. If the amount of sequence identity is significant, then the two sequences are probably homologous. It is never correct to say that two proteins share a certain percent homology, because they are either homologous or not. We will discuss the statistical significance of sequence alignments below; these analyses provide evidence to assess the hypothesis that two proteins are homologous. Ultimately the strongest evidence to determine whether two proteins are homologous comes from structural studies in combination with evolutionary analyses.

Two proteins could have similar structures due to convergent evolution. Thus molecular evolutionary studies are essential.

Gaps

Pairwise alignment is useful as a way to identify mutations that have occurred during evolution and have caused divergence of the sequences of the two proteins we are studying. The most common mutations are *substitutions, insertions,* and *deletions.* Substitutions occur when a mutation results in the codon for one amino acid being changed into that for another. This results in the alignment of two nonidentical amino acids, such as serine and threonine. Insertions and deletions occur when residues are added or removed and are represented by null characters (the period symbol, .) that are added to one or the other sequence. Insertions or deletions (even rhose just one character long) are referred to as *gaps* in the alignment.

In our alignment of human RBP4 and bovine β-lactoglobulin there are eight gaps. Two occur at the ends of the proteins, two long gaps are in the middle of the alignment, and there are four gaps of just one or two amino acids. Note that one of the effects of the gaps is to make the overall length of each alignment exactly the same.

The addition of gaps in an alignment may be biologically relevant because the gaps reflect evolutionary changes that have occurred. Practically, gaps can allow the full alignment of two proteins.

Pairwise Alignment, Homology, and Evolution of Life

If two proteins are homologous, they share a common ancestor. Generally, we observe the sequence of proteins (and genes) from organisms that are extant. We can compare RBP from species such as human and a fish, rainbow trout, and see that the sequences are homologous (Fig. 3.2). This implies that an ancestral organism had

It is possible to infer the sequence of the common ancestor (see Chapter 11).

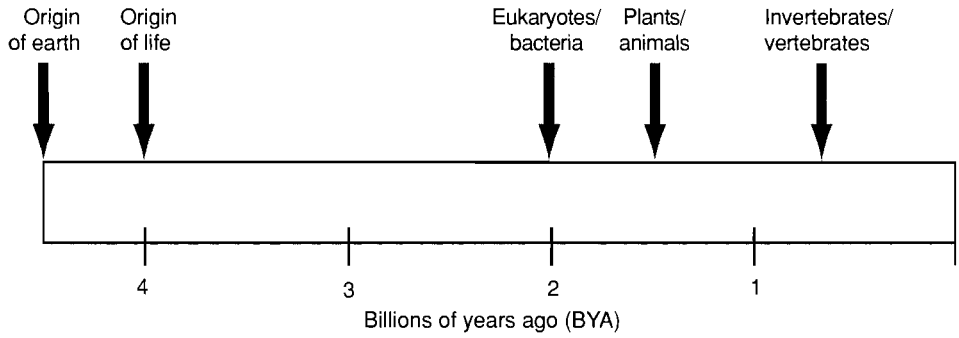


FIGURE 3.6. *Overview of the history of life on Earth. See Chapter 12 for details. Gene/protein sequences are analyzed in the context of evolution: Which organisms have orthologous genes? When did these organisms evolve? How related are human and bacterial lipocalins?*

an RBP gene and lived sometime before the divergences of the lineages that gave rise to human and trout (over 400 MYA). Descendants of that ancestral organism include many species in addition to human and trout. The study of homologous protein (or DNA) sequences by pairwise alignment involves an investigation of the evolutionary history of that protein (or gene).

For a brief overview of the time scale of life on Earth, see Figure 3.6 (refer to Chapter 12 for a more detailed discussion). The divergence of different species is established through the use of many sources of data, especially the fossil record. Fossils of prokaryotes have been discovered in rocks 3.5 billion years old. In the case of lipocalins, no invertebrate (e.g., insect) ortholog of RBP has been identified, but several fish and amphibian RBPs are known (Fig. 3.2). We can therefore infer that the RBP gene originated between 700 and 400 MYA. Other lipocalin genes are more ancient. For example, bacterial lipocalin genes presumably arose 2 BYA.

As we examine a variety of homologous protein sequences, we can observe a wide range of conservation between family members. Some are very ancient and well conserved, such as the enzyme glyceraldehyde-3-phosphate dehydrogenase (GAPDH).

```
fly        GAKKVIISAP SAD.APM..F VCGVNLDAYK PDMKVVSNAS CTTNCLAPLA
human      GAKRVIISAP SAD.APM..F VMGVNHEKYD NSLKIISNAS CTTNCLAPLA
plant      GAKKVIISAP SAD.APM..F VVGVNEHTYQ PNMDIVSNAS CTTNCLAPLA
bacterium  GAKKVVMTGP SKDNTPM..F VKGANFDKY. AGQDIVSNAS CTTNCLAPLA
yeast      GAKKVVITAP SS.TAPM..F VMGVNEEKYT SDLKIVSNAS CTTNCLAPLA
archaeon   GADKVLISAP PKGDEPVKQL VYGVNHDEYD GE.DVVSNAS CTTNSITPVA

fly        KVINDNFEIV EGLMTTVHAT TATQKTVDGP SGKLWRDGRG AAQNIIPAST
human      KVIHDNFGIV EGLMTTVHAI TATQKTVDGP SGKLWRDGRG ALQNIIPAST
plant      KVVHEEFGIL EGLMTTVHAT TATQKTVDGP SMKDWRGGRG ASQNIIPSST
bacterium  KVINDNFGII EGLMTTVHAT TATQKTVDGP SHKDWRGGRG ASQNIIPSST
yeast      KVINDAFGIE EGLMTTVHSL TATQKTVDGP SHKDWRGGRT ASGNIIPSST
archaeon   KVLDEEFGIN AGQLTTVHAY TGSQNLMDGP NGKP.RRRRA AAENIIPTST

fly        GAAKAVGKVI PALNGKLTGM AFRVPTPNVS VVDLTVRLGK GASYDEIKAK
human      GAAKAVGKVI PELNGKLTGM AFRVPTANVS VVDLTCRLEK PAKYDDIKKV
plant      GAAKAVGKVL PELNGKLTGM AFRVPTSNVS VVDLTCRLEK GASYEDVKAA
bacterium  GAAKAVGKVL PELNGKLTGM AFRVPTPNVS VVDLTVRLEK AATYEQIKAA
yeast      GAAKAVGKVL PELQGKLTGM AFRVPTVDVS VVDLTVKLNK ETTYDEIKKV
archaeon   GAAQAATEVL PELEGKLDGM AIRVPVPNGS ITEFVVDLDD DVTESDVNAA
```

FIGURE 3.7. *Multiple sequence alignment of a portion of the glyceraldehyde 3-phosphate dehydrogenase (GAPDH) protein from six organisms:* Drosophila melanogaster *(the fly; GenBank accession P07487),* H. sapiens *(NP_002037),* Arabidopsis thaliana *(a plant; AAD30223), the bacterium* Escherichia coli *(P06977),* Saccharomyces cerevisiae *(yeast; NP_011708), and an archaeon,* Haloarcula vallismortis *(Q48335). The alignment was performed using the PileUp program of GCG (see Chapter 10 and Appendix). Note that the archaeal sequence is most divergent, but overall these orthologs are extremely well conserved. (Archaea are microbes that form one of the three branches of life; see Chapters 12 and 14.)*

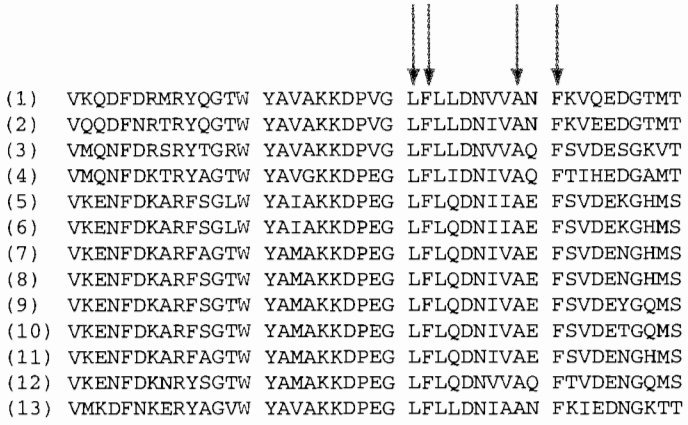


FIGURE 3.8. *Partial multiple sequence alignment of RBPs. Note that many columns of residues are perfectly conserved (including the GXW motif, shaded). Accession numbers are given in legend to Figure 3.2. The proteins are from (1) common carp,* C. carpio; *(2) zebrafish,* D. rerio; *(3) rainbow trout,* O. mykiss; *(4) hermaphrodite,* S. aurata; *(5) mouse,* M. musculus; *(6) rat,* R. norvegicus; *(7) cow,* B. taurus; *(8) pig,* S. scrofa; *(9) horse,* E. caballa; *(10) human,* H. sapiens; *(11) rabbit,* O. cuniculus; *(12) chicken,* G. gallus; *and (13) frog,* X. laevis.

A multiple sequence alignment, which is essentially a series of pairwise alignments between a group of proteins, reveals that GAPDH orthologs are extraordinarily well conserved (Fig. 3.7).

Orthologous RBPs from various species provide another example of a well-conserved family (Fig. 3.8). Many columns in this alignment are perfectly conserved, including the glycine–X–tryptophan (GXW; X refers to any amino acid) motif that is characteristic of hundreds of lipocalin proteins. Notice that some positions include amino acids that represent conservative substitutions, such as the choice of serine or threonine in the final position of the alignment (Fig. 3.8). Other positions are less well conserved; immediately preceding the canonical GXW motif, the amino acid may be glutamine, threonine, serine, or alanine. Some of the amino acid residues that form a binding pocket for retinol are perfectly conserved (Newcomer et al., 1984) (Fig. 3.8, arrows).

The human lipocalin family provides a final example of a multiple sequence alignment. Members of this family are paralogous but highly divergent (Fig. 3.9). The only well-conserved motif in this alignment is the canonical GXW signature.

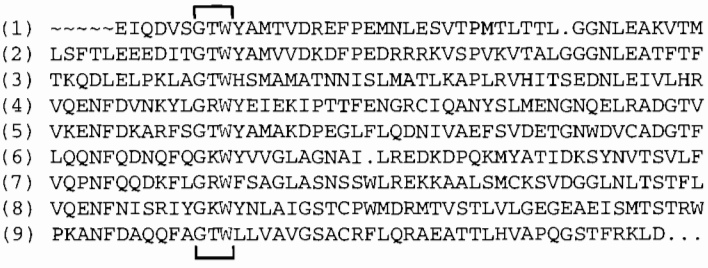


FIGURE 3.9. *Partial multiple sequence alignment of human lipocalin paralogs. Note the conserved GXW motif (brackets). All sequences are* H. sapiens. *The proteins are (1) lipocalin 1, (2) odorant-binding protein 2A, (3) progestagen-associated endometrial protein, (4) apolipoprotein D, (5)* RBP4, *(6) neutrophil gelatinase-associated protein, (7) prostaglandin D2 synthase, (8) α-1-microglobulin/bikunin, and (9) complement component 8, gamma peptide. The accession numbers are given in the legend to Figure 3.3. Note that many columns of this alignment have been removed where there were gaps in the sequences.*

Despite the tremendous divergence of the amino acid sequences, it is likely that all members of this family adopt a highly similar three-dimensional structure (see Fig. 3.1 and Chapter 9).

We can see from the preceding examples that pairwise sequence alignment between any two proteins can exhibit widely varying amounts of conservation. Margaret Dayhoff (1978) provided a model of the rules by which evolutionary change occurs in proteins. We will now examine the Dayhoff model, which provides the basis of a quantitative scoring system for pairwise alignments. This system accounts for scores between any proteins, whether they are closely or distantly related. Then we will discuss the two main kinds of pairwise sequence algorithms, global and local. Many database searching methods such as BLAST (Chapters 4 and 5) depend in some form upon the evolutionary insights of the Dayhoff model.

The Dayhoff (1978) reference is to the *Atlas of Protein Sequence and Structure,* a book with 25 chapters (and various coauthors) describing protein families.

DAYHOFF MODEL: ACCEPTED POINT MUTATIONS

Dayhoff and colleagues catalogued thousands of proteins and compared the sequences of closely related proteins in many families. They considered the question of which specific amino acid substitutions are observed to occur when two homologous protein sequences are aligned. They defined an *accepted point mutation* as a replacement of one amino acid in a protein by another residue that has been accepted by natural selection. Accepted point mutation is abbreviated *PAM* (which is easier to pronounce than APM). An amino acid change that is accepted by natural selection

Dayhoff and colleagues did not compare the probability of one residue mutating directly into another. Instead, they constructed phylogenetic trees using parsimony analysis (see Chapter 11). Then, they described the probability that two aligned residues derived from a common ancestral residue. With this approach, they could minimize the confounding effects of multiple substitutions occurring in an aligned pair of residues.

BOX 3-3
Dayhoff's Protein Superfamilies

Dayhoff (1978, p. 3) studied 34 protein "superfamilies" grouped into 71 phylogenetic trees. These proteins ranged from some that are very well conserved (e.g., histones and glutamate dehydrogenase) to others that have a high rate of mutation acceptance [e.g., immunoglobulin (Ig) chains and carrier proteins]. Protein families were aligned (compare Fig. 3.7); then they counted how often any one amino acid in the alignment was replaced by another. Here is a partial list of the proteins they studied, including the rates of mutation acceptance. For a more detailed list, see Table 11.1 below. There is a range of almost 400-fold between the families that evolve fastest and slowest, but within a given family the rate of evolution (measured in PAMs per unit time) varies only two- to threefold between species. Used with permission.

Protein	PAMs per 100 million years
Immunoglobulin (Ig) kappa chain C region	37
Kappa casein	33
Epidermal growth factor	26
Serum albumin	19
Hemoglobin alpha chain	12
Myoglobin	8.9
Nerve growth factor	8.5
Trypsin	5.9
Insulin	4.4
Cytochrome *c*	2.2
Glutamate dehydrogenase	0.9
Histone H3	0.14
Histone H4	0.10

occurs when (1) a gene undergoes a DNA mutation such that it encodes a different amino acid and (2) the entire species adopts that change as the predominant form of the protein.

Which point mutations are accepted in protein evolution? Intuitively, conservative replacements such as serine for threonine would be most readily accepted. In order to determine all possible changes, Dayhoff and colleagues examined 1572 changes in 71 groups of closely related proteins (Box 3.3). Thus their definition of "accepted" mutations was based on empirically observed amino acid changes. The results are shown in Figure 3.10, which describes the frequency with which any amino acid pairs *i,j* are aligned. Today, we could generate a table like this with vastly more data (refer to Fig. 2.1 and the explosive growth of GenBank). Gonnet and others have produced updated versions of the PAM matrices (Gonnet et al., 1992). Nonetheless the findings from 1978 are essentially correct. Some amino acid residues such as asparagine and serine undergo substitutions very frequently, while other residues (notably tryptophan and cysteine) are mutable only very rarely.

Dayhoff et al. calculated the relative mutabilities of the amino acids (Table 3.1). This simply describes how often each amino acid is likely to change over a short evolutionary period. (We note that the evolutionary period in question is short because this analysis involves protein sequences that are closely related to each other.) To calculate relative mutability, they divided the number of times each amino acid was observed to mutate by the overall frequency of occurrence of that amino acid. Table 3.2 shows the frequency with which each amino acid is found.

Why are some amino acids more mutable than others? The less mutable residues probably have important structural or functional roles in proteins, such that the consequence of replacing them with any other residue could be harmful to the organism. (We will see in Chapter 18 that many human diseases, from cystic fibrosis to the autism-related Rett syndrome, can be caused by a single amino acid substitution in a protein.) Conversely, the most mutable amino acids—asparagine, serine, aspartic acid, and glutamic acid—have functions in proteins that are easily assumed by other residues. The most common substitutions seen in Figure 3.10 are glutamic acid for aspartic acid (both are acidic), serine for alanine, serine for threonine (both are hydroxylated), and isoleucine for valine (both are hydrophobic and of a similar size).

The substitutions that occur in proteins can also be understood with reference to the genetic code (Box 3.4). Observe how common amino acid substitutions tend to require only a single nucleotide change. For example, aspartate is encoded by GAU or GAC, and changing the third position to either A or G causes the codon to encode a glutamic acid. Also note that four of the five least mutable amino acids (tryptophan, cysteine, phenylalanine, and tyrosine) are specified by only one or two codons. A mutation of any of the three bases of the W codon is guaranteed to change that amino acid. The low mutability of this amino acid suggests that substitutions are not tolerated by natural selection. Of the eight least mutable amino acids (Table 3.1), only one (leucine) is specified by six codons, and only two (glycine and proline) are specified by four codons. The others are specified by one or two codons.

We will consider protein superfamilies in Chapter 8.

Dayhoff et al. focused on proteins sharing 85% or more identity. Thus, they could construct their alignments with a high degree of confidence. Later in this chapter, we will see how the Needleman and Wunsch algorithm (described in 1970) permits the automated alignment of protein sequences.

The observed frequencies of substitution for each pair of amino acids are called the "target frequencies." They form the basis of the substitution matrices that we describe in this section.

You can look up the frequency of occurrence of each amino acid using the proteome tools at the European Bioinformatics Institute (EBI) at ▸http://www.ebi.ac.uk/proteome/. Go there, click *Homo sapiens,* then click "primary amino acid composition."

PAM1 Matrix

Dayhoff and colleagues next used the data on accepted mutations (Fig. 3.10) and the probabilities of occurrence of each amino acid to generate a *mutation probability matrix M* (Fig. 3.11). Each element of the matrix M_{ij} shows the probability that an original amino acid j (see the columns) will be replaced by another amino acid i (see

Original amino acid

Replacement amino acid	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	0	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	
	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val

FIGURE 3.10. *Numbers of accepted point mutations, multiplied by 10, in 1572 cases of amino acid substitutions from closely related protein sequences. This figure is modified from Dayhoff (1978, p. 346). Amino acids are presented alphabetically according to the three-letter code. Notice that some substitutions are very commonly accepted (such as V and I or S and T). Other amino acids, such as C and W, are rarely substituted by any other residue. Used with permission.*

the rows) over a defined evolutionary interval. In the case of Fig. 3.11 the interval is one PAM, which is defined as the unit of evolutionary divergence in which 1% of the amino acids have been changed between the two protein sequences. Note that the evolutionary interval of this PAM matrix is defined in terms of percent amino acid divergence and not in units of years. As we have seen (compare Figs. 3.7, 3.8, and 3.9), 1% divergence of protein sequence may occur over vastly different time frames for protein families that undergo substitutions at different rates.

Examination of Figure 3.11 reveals several important features. The highest scores are distributed in a diagonal from top left to bottom right. The values in each column sum to 10,000 (this corresponds to 100%). The value 9867 at the top left indicates that when the original sequence consists of an alanine there is a 98.67% chance that the replacement amino acid will also be an alanine over an evolutionary distance of one PAM. There is a 0.28% chance that it will be changed to serine. The most mutable amino acid (from Table 3.1), asparagine, has only a 98.22% chance of remaining unchanged; the least mutable amino acid, tryptophan, has a 99.7% chance of remaining the same.

For each original amino acid, it is easy to observe the amino acids that are most likely to replace it if a change should occur. These data are very relevant to pairwise sequence alignment because they will form the basis of a scoring system (described below) in which reasonable amino acid substitutions in an alignment are rewarded while unlikely substitutions are penalized. These concepts are also relevant to database searching algorithms such as BLAST (Chapters 4 and 5) which depend upon rules to score the relatedness of molecular sequences.

Almost all molecular sequence data are obtained from extant organisms. We can infer ancestral sequences, as will be described in Chapter 11. But in general, for an aligned pair of residues a, b we do not know which one mutated into the other. Dayhoff and colleagues used the assumption that accepted amino acid mutations are undirected, that is, they are equally likely in either direction. In the PAM1 matrix, the close relationship of the proteins makes it unlikely that the ancestral residue is entirely different than both of the observed, aligned residues.

TABLE 3-1 Relative Mutabilities of Amino Acids

The value of alanine is arbitrarily set to 100.

Asn	134	His	66
Ser	120	Arg	65
Asp	106	Lys	56
Glu	102	Pro	56
Ala	100	Gly	49
Thr	97	Tyr	41
Ile	96	Phe	41
Met	94	Leu	40
Gln	93	Cys	20
Val	74	Trp	18

Source: From Dayhoff (1978). Used with permission.

PAM250 AND OTHER PAM MATRICES

The PAM1 matrix was based upon the alignment of closely related protein sequences, all of which were at least 85% identical within a protein family. We are often interested in exploring the relationships of proteins that share far less than 85% amino acid identity. We can accomplish this by constructing probability matrices for proteins that share any degree of amino acid identity. Consider closely related proteins, such as the GAPDH proteins shown in Figure 3.7. A mutation from one residue to another is a relatively rare event, and a scoring system used to align two such closely related proteins should reflect this. In the PAM1 matrix (Fig. 3.11) some substutions such as tryptophan to threonine are so rare that they are never observed in the data set. But next consider two distantly related proteins, such as the lipocalins shown in Figure 3.9. Here, substitutions are likely to be very common. PAM matrices such as PAM100 or PAM250 were generated to reflect the kinds of amino acid substitutions that occur in distantly related proteins.

How are PAM matrices other than PAM1 derived? Dayhoff et al. multiplied the PAM1 matrix by itself, up to hundreds of times, to obtain other PAM matrices (see Box 3.5). Thus they extrapolated from the PAM1 matrix.

TABLE 3-2 Normalized Frequencies of Amino Acids

These values sum to 1. If the 20 amino acids were equally represented in proteins, these values would all be 0.05 (i.e. 5%); instead, amino acids vary in their frequency of occurrence

Gly	0.089	Arg	0.041
Ala	0.087	Asn	0.040
Leu	0.085	Phe	0.040
Lys	0.081	Gln	0.038
Ser	0.070	Ile	0.037
Val	0.065	His	0.034
Thr	0.058	Cys	0.033
Pro	0.051	Tyr	0.030
Glu	0.050	Met	0.015
Asp	0.047	Trp	0.010

Source: From Dayhoff (1978). Used with permission.

BOX 3-4
Genetic Code

UUU	171	F	UCU	147	S	UAU	124	Y	UGU	99	C
UUC	203	F	UCC	172	S	UAC	158	Y	UGC	119	C
UUA	73	L	UCA	118	S	UAA	0	Stop	UGA	0	Stop
UUG	125	L	UCG	45	S	UAG	0	Stop	UGG	122	W
CUU	127	L	CCU	175	P	CAU	104	H	CGU	47	R
CUC	187	L	CCC	197	P	CAC	147	H	CGC	107	R
CUA	69	L	CCA	170	P	CAA	121	Q	CGA	63	R
CUG	392	L	CCG	69	P	CAG	343	Q	CGG	115	R
AUU	165	I	ACU	131	T	AAU	174	N	AGU	121	S
AUC	218	I	ACC	192	T	AAC	199	N	AGC	191	S
AUA	71	I	ACA	150	T	AAA	248	K	AGA	113	R
AUG	221	M	ACG	63	T	AAG	331	K	AGG	110	R
GUU	111	V	GCU	185	A	GAU	230	D	GGU	112	G
GUC	146	V	GCC	282	A	GAC	262	D	GGC	230	G
GUA	72	V	GCA	160	A	GAA	301	E	GGA	168	G
GUG	288	V	GCG	74	A	GAG	404	E	GGG	160	G

In this table, the 64 possible codons are depicted along with the frequency of codon utilization and the single-letter code of the amino acid that is specified. There are four bases (A, C, G, U) and three bases per codon, so there are $4 \times 4 \times 4 = 4^3 = 64$ codons.

Several features of the genetic code should be noted. Amino acids may be specified by one codon (M, W), two codons (C, D, E, F, H, K, N, Q, Y), three codons (I), four codons (A, G, P, T, V), or six codons (L, R, S). UGA is rarely read as a selenocysteine (abbreviated sec, and the assigned single-letter abbreviations is U).

For each block of four codons that are grouped together, one is often used dramatically less frequently. For example, for F, L, I, M, and V (i.e. codons with a U in the middle, occupying the first column of the genetic code), adenine is used relatively infrequently in the third-codon position. For codons with a cytosine in the center, guanine is strongly underrepresented in the third position.

Also note that in many cases mutations cause a conservative change (or no change at all) in the amino acid. Consider threonine (ACX). Any mutation in the third position causes no change in the specified amino acid, because of "wobble." If the first nucleotide of any threonine codon is mutated from A to U, the conservative replacement to a serine occurs. If the second nucleotide C is mutated to a G, a serine replacement occurs. Similar patterns of conservative substitution can be seen along the entire first column of the genetic code, where all of the residues are hydrophobic, and for the charged residues D, E and K, R as well.

Codon usage varies between organisms and between genes within organisms. Note also that while this is the standard genetic code, some organisms use alternate genetic codes.

Source: Adapted from the International Human Genome Sequencing Consortium (2001), Figure 34. Used with permission.

To make sense of what different PAM matrices mean, consider the extreme cases. When PAM equals zero, the matrix is a unit diagonal (Fig. 3.12), because no amino acids have changed. PAM can be extremely large (e.g., PAM greater than 2000, or the matrix can even be multiplied against itself an infinite number of times). In the resulting PAM∞ matrix there is an equal likelihood of any amino acid

Original amino acid

Replacement amino acid

	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	2	2	3	1	2	1	2	0	0	9872	9	2	21	7	0	1	7	0	1	33
L	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

FIGURE 3.11. *The PAM1 mutation probability matrix. From Dayhoff (1978, p. 348, Fig. 82). The original amino acid j is arranged in columns (across the top), while the replacement amino acid i is arranged in rows. Used with permission.*

BOX 3-5
Matrix Multiplication

A matrix is an orderly array of numbers. An example of a matrix with rows i and columns j is:

$$\begin{bmatrix} 1 & 2 & 4 \\ 2 & 0 & -3 \\ 4 & -3 & 6 \end{bmatrix}$$

In a symmetric matrix, such as the one above, $a_{ij} = a_{ji}$. This means that all the corresponding nondiagonal elements are equal. Matrices may be added, subtracted, or manipulated in a variety of ways. Two matrices can be multipled together provided that the number of columns in the first matrix M_1 equals the number of rows in the second matrix M_2. Following is an example of how to multiply M_1 by M_2.

Successively multiply each row of M_1 by each column of M_2:

$$M_1 = \begin{bmatrix} 3 & 4 \\ 0 & 2 \end{bmatrix} \quad M_2 = \begin{bmatrix} 5 & -2 \\ 2 & 1 \end{bmatrix}$$

$$M_{12} = \begin{bmatrix} (3)(5)+(4)(2) & (3)(-2)+(4)(1) \\ (0)(5)+(2)(2) & (0)(-2)+(2)(1) \end{bmatrix} = \begin{bmatrix} 23 & -2 \\ 4 & 2 \end{bmatrix}$$

If you want to try matrix multiplication yourself, enter the PAM1 mutation probability matrix of Figure 3.11 into a program such as Matlab (Mathworks), divide each value by 10,000, and multiply the matrix times itself 250 times. You will get the PAM250 matrix of Figure 3.13.

being present and all the values consist of rows of probabilities that approximate the background probability for the frequency occurrence of each amino acid (Fig. 3.12, lower panel). We described these background frequencies in Table 3.2.

The PAM250 matrix is of particular interest (Fig. 3.13). It is produced when the PAM1 matrix is multiplied against itself 250 times, and it is one of the common matrices used for BLAST searches of databases (Chapter 4). This matrix applies to

Original amino acid

Replacement amino acid

PAM0	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly
A	100%	0%	0%	0%	0%	0%	0%	0%
R	0%	100%	0%	0%	0%	0%	0%	0%
N	0%	0%	100%	0%	0%	0%	0%	0%
D	0%	0%	0%	100%	0%	0%	0%	0%
C	0%	0%	0%	0%	100%	0%	0%	0%
Q	0%	0%	0%	0%	0%	100%	0%	0%
E	0%	0%	0%	0%	0%	0%	100%	0%
G	0%	0%	0%	0%	0%	0%	0%	100%

Original amino acid

Replacement amino acid

PAM∞	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly
A	8.7%	8.7%	8.7%	8.7%	8.7%	8.7%	8.7%	8.7%
R	4.1%	4.1%	4.1%	4.1%	4.1%	4.1%	4.1%	4.1%
N	4.0%	4.0%	4.0%	4.0%	4.0%	4.0%	4.0%	4.0%
D	4.7%	4.7%	4.7%	4.7%	4.7%	4.7%	4.7%	4.7%
C	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%	3.3%
Q	3.8%	3.8%	3.8%	3.8%	3.8%	3.8%	3.8%	3.8%
E	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%	5.0%
G	8.9%	8.9%	8.9%	8.9%	8.9%	8.9%	8.9%	8.9%

FIGURE 3.12. *Portion of the matrices for a zero PAM value (PAM0; upper panel) or for an infinite PAM∞ value (lower panel). At PAM∞ (i.e., if the PAM1 matrix is multiplied against itself an infinite number of times), all the entries in each row converge on the normalized frequency of the replacement amino acid (see Table 3.2). A PAM2000 matrix has similar values that tend to converge on these same limits. In a PAM2000 matrix, the proteins being compared are at an extreme of unrelatedness. In contrast, at PAM0, no mutations are tolerated, and the residues of the proteins are perfectly conserved.*

Original amino acid

Replacement amino acid

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

FIGURE 3.13. *The PAM250 mutation probability matrix. From Dayhoff (1978, p. 350, Fig. 83). At this evolutionary distance, only one in five amino acid residues remains unchanged from an original amino acid sequence (columns) to a replacement amino acid (rows). Note that the scale has changed relative to Fig. 3.11, and the columns sum to 100. Used with permission.*

an evolutionary distance where proteins share about 20% amino acid identity. Compare this matrix to the PAM1 matrix (Fig. 3.11) and note that much of the information content is lost. The diagonal from top left to bottom right tends to contain higher values than elsewhere in the matrix, but not in the dramatic fashion of the PAM1 matrix. As an example of how to read the PAM250 matrix, if the original amino acid is an alanine, there is just a 13% chance that the second sequence will also have an alanine. In fact, there is a nearly equal probability (12%) that the alanine will have been replaced by a glycine. For the least mutable amino acids, tryptophan and cysteine, there is more than a 50% probability that those residues will remain unchanged at this evolutionary distance.

From a Mutation Probability Matrix to a Log-Odds Score Matrix

Our goal in studying PAM matrices is to derive a scoring system so that we can assess the relatedness of two sequences. When we perform BLAST searches (Chapters 4 and 5) or pairwise alignments, we employ a scoring matrix, but it is not in the form we have described so far. The PAM250 mutation probability matrix (e.g., Fig. 3.13) is useful because it describes the frequency of amino acid replacements between distantly related proteins. We next need to convert the elements of a PAM mutation probability matrix into a scoring matrix, also called a log-odds matrix or relatedness odds matrix.

The cells in a log-odds matrix consist of an "odds ratio" between two probabilities that describe the probability that some amino acid a will change to amino acid b in some PAM interval. The score S for an alignment of a, b is given by

$$S\,(\mathrm{a}, \mathrm{b}) = 10\,\log_{10}(M_{\mathrm{ab}}/p_{\mathrm{b}}) \qquad (3.1)$$

In Equation 3.1, the two probabilities that form the "odds ratio" are as follows. (1) The term M_{ab} (e.g., from Fig. 3.13 in the particular case of a PAM250 matrix) is the

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	2																			
R	-2	6																		
N	0	0	2																	
D	0	-1	2	4																
C	-2	-4	-4	-5	12															
Q	0	1	1	2	-5	4														
E	0	-1	1	3	-5	2	4													
G	1	-3	0	1	-3	-1	0	5												
H	-1	2	2	1	-3	3	1	-2	6											
I	-1	-2	-2	-2	-2	-2	-2	-3	-2	5										
L	-2	-3	-3	-4	-6	-2	-3	-4	-2	-2	6									
K	-1	3	1	0	-5	1	0	-2	0	-2	-3	5								
M	-1	0	-2	-3	-5	-1	-2	-3	-2	2	4	0	6							
F	-3	-4	-3	-6	-4	-5	-5	-5	-2	1	2	-5	0	9						
P	1	0	0	-1	-3	0	-1	0	0	-2	-3	-1	-2	-5	6					
S	1	0	1	0	0	-1	0	1	-1	-1	-3	0	-2	-3	1	2				
T	1	-1	0	0	-2	-1	0	0	-1	0	-2	0	-1	-3	0	1	3			
W	-6	2	-4	-7	-8	-5	-7	-7	-3	-5	-2	-3	-4	0	-6	-2	-5	17		
Y	-3	-4	-2	-4	0	-4	-4	-5	0	-1	-1	-4	-2	7	-5	-3	-3	0	10	
V	0	-2	-2	-2	-2	-2	-2	-1	-2	4	2	-2	2	-1	-1	-1	0	-6	-2	4

FIGURE 3.14. *Log-odds matrix for PAM250. High PAM values (e.g., PAM250) are useful for aligning very divergent sequences. A variety of algorithms for pairwise alignment, multiple sequence alignment, and database searching (e.g., BLAST) allow you to select an assortment of PAM matrices such as PAM250, PAM70, and PAM30.*

probability that the aligned pair of amino acid residues a, b represents an authentic alignment (i.e., a mutation accepted by evolution). (2) The normalized frequency p_b represents the probability that the residue b was aligned by random chance. This second term reflects the independent probabilities of each amino acid a, b occurring in this position. Its values were given in Table 3.2. The likelihood ratio, or odds ratio, is then M_{ab}/p_b.

For this scoring system Dayhoff and colleagues took 10 times the base 10 logarithm of the odds ratio (see Equation 3.1). Using the logarithm here is helpful because it allows us to sum the scores of the aligned residues when we perform an overall alignment of two sequences. (If we did not take the logarithm, we would need to multiply the ratios at all the aligned positions, and this is computationally more cumbersome.)

A log-odds matrix for PAM250 is shown in Figure 3.14. The values have been rounded off to the nearest integer. Try using Equation 3.1 to make sure you understand how the mutation probability matrix (Fig. 3.13) is converted into the log-odds matrix (Fig. 3.14). As an example, for tryptophan the PAM250 mutation probability matrix value is 0.55, and the normalized frequency of tryptophan is 0.010. Thus,

$$S(\text{a, tryptophan}) = 10 \log_{10}(0.55/0.010) = 17.4 \qquad (3.2)$$

This value is rounded off to 17 in the PAM250 log-odds matrix. What do the scores in the PAM250 matrix signify? A score of −10 indicates that the correspondence of two amino acids in an alignment that accurately represents homology (evolutionary descent) is one-tenth as frequent as the chance alignment of these amino acids. This assumes that each was randomly selected from the background amino acid frequency distribution. A score of zero is neutral. A score of +17 for tryptophan indicates that this correspondence is 50 times more frequent than the chance alignment of this residue in a pairwise alignment. A score of +2 indicates that the amino acid replacement occurs 1.6 times as frequently as expected by chance. The highest values in this particular matrix are for tryptophan (17 for an identity) and cysteine (12), while the most severe penalties are associated with substitutions for those two residues. When two sequences are aligned and a score is given, that score is simply the sum of the scores for all the aligned residues across the alignment.

Note that this scoring matrix is symmetric, in contrast to the mutation probability matrix in Figure 3.13. In a comparison of two sequences it does not matter which is given first.

We state that a score of +17 for tryptophan indicates that the correspondence of two tryptophans in an alignment of homologous proteins is 50 times more likely than a chance alignment of two tryptophan residues. How do we derive the number 50? From Equation 3.1, let S (a, b) = +17 and let the probability of replacement $M_{ab}/p_b = x$. Then $+17 = 10 \log_{10} x$, $+1.7 = \log_{10} x$, and $10^{1.7} = x = 50$.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	7																			
R	-10	9																		
N	-7	-9	9																	
D	-6	-17	-1	8																
C	-10	-11	-17	-21	10															
Q	-7	-4	-7	-6	-20	9														
E	-5	-15	-5	0	-20	-1	8													
G	-4	-13	-6	-6	-13	-10	-7	7												
H	-11	-4	-2	-7	-10	-2	-9	-13	10											
I	-8	-8	-8	-11	-9	-11	-8	-17	-13	9										
L	-9	-12	-10	-19	-21	-8	-13	-14	-9	-4	7									
K	-10	-2	-4	-8	-20	-6	-7	-10	-10	-9	-11	7								
M	-8	-7	-15	-17	-20	-7	-10	-12	-17	-3	-2	-4	12							
F	-12	-12	-12	-21	-19	-19	-20	-12	-9	-5	-5	-20	-7	9						
P	-4	-7	-9	-12	-11	-6	-9	-10	-7	-12	-10	-10	-11	-13	8					
S	-3	-6	-2	-7	-6	-8	-7	-4	-9	-10	-12	-7	-8	-9	-4	7				
T	-3	-10	-5	-8	-11	-9	-9	-10	-11	-5	-10	-6	-7	-12	-7	-2	8			
W	-20	-5	-11	-21	-22	-19	-23	-21	-10	-20	-9	-18	-19	-7	-20	-8	-19	13		
Y	-11	-14	-7	-17	-7	-18	-11	-20	-6	-9	-10	-12	-17	-1	-20	-10	-9	-8	10	
V	-5	-11	-12	-11	-9	-10	-10	-9	-9	-1	-5	-13	-4	-12	-9	-10	-6	-22	-10	8

FIGURE 3.15. *Log-odds matrix for PAM10. Low PAM values such as this are useful for aligning very closely related sequences. Compare this with the PAM250 matrix (Fig. 3.14) and note that there are larger positive scores for identical matches in this PAM10 matrix and larger penalties for mismatches.*

It is easy to see how different PAM matrices score amino acid substitutions by comparing the PAM250 matrix (Fig. 3.14) with a PAM10 matrix (Fig. 3.15). In the PAM10 matrix, identical amino acid residue pairs tend to produce a higher score than in the PAM250 matrix; for example, a match of alanine to alanine scores 7 versus 2, respectively. The penalties for mismatches are greater in the PAM10 matrix; for example, a mutation of aspartate to arginine scores −17 (PAM10) versus −1 (PAM250). PAM10 even has negative scores for substitutions (such as glutamate to asparagine, −5) that are scored positively in the PAM250 matrix (+1). To see the relevance of PAM250 and PAM10 matrices, consider an alignment of two distantly related lipocalins. Using the PAM250 matrix, many amino acid substitutions would receive relatively minor penalties, and a significant score is possible. But using the PAM10 matrix, amino acid substitutions would earn large penalties and would generally not be tolerated. For two closely related lipocalins sharing a large amount of amino acid identity (e.g., mouse and rat RBP), a PAM10 matrix would be more appropriate for the evolutionary distance of those proteins.

Practical Usefulness of PAM Matrices in Pairwise Alignment

We will demonstrate the usefulness of PAM matrices by performing a pairwise alignment of two proteins and examining the outcome using the PAM40 versus the PAM250 matrix. Try this using the web-based SIM alignment program. The ExPASy site provides two boxes; paste the sequences (or alternatively paste in the protein accession numbers) of human RBP4 and bovine β-lactoglobulin. Note that you can select from different PAM matrices using a pull-down menu ("comparison matrix"). First try the PAM250 matrix; this matrix is appropriate to use because the two proteins are only distantly related. An overlap of 20 identical residues over a span of

The ExPASy website (http://www.expasy.ch) includes the SIM alignment program (http://www.expasy.ch/tools/sim-prot.html).

(*a*)

```
24.7% identity in 81 residues overlap; Score: 77.0; Gap frequency: 3.7%
 hsrbp,  26 RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDV
btlact,  21 QTMKGLDIQKVAGTWYSLAMAASD-ISLLDAQSAPLRVYVEELKPTPEGDLEILLQKWEN
                 *     ****  *        * *       *                 **  *

hsrbp,   86 --CADMVGTFTDTEDPAKFKM
btlact,  80 GECAQKKIIAEKTKIPAVFKI
              **        *  ** **
```

(*b*)

```
60.0% identity in 10 residues overlap; Score: 23.0; Gap frequency: 0.0%
hsrbp,   136 CRLLNLDGTC
btlact,    3 CLLLALALTC
             * ** *  **
```

FIGURE 3.16. *Pairwise alignment of RBP4 and β-lactoglobulin using (a) a PAM250 matrix or (b) a PAM40 matrix. The PAM250 matrix is better able to detect regions of similarity between distantly related proteins. Note that the PAM40 matrix resulted in an inappropriate alignment between a carboxy-terminal portion of RBP4 (beginning at amino acid 136) and an amino-terminal portion of β-lactoglobulin. The asterisks indicate positions of amino acid identity.*

81 amino acids is detected (24.7% identity) (Fig. 3.16*a*). Next try a PAM40 matrix. Now, the best-aligned segment is only 10 amino acids in length (Fig. 3.16*b*). Moreover, this short alignment is biologically meaningless. The PAM40 matrix is not appropriate for detecting distantly related protein sequences. An important concept is that different scoring matrices vary in their sensitivity to protein sequences (or DNA sequences) of varying relatedness.

Other pairwise sequence alignment programs are described later (see Tables 3.4 and 3.5 under Web Resources).

Important Alternative to PAM: BLOSUM Scoring Matrices

In addition to the PAM matrices, another very common set of scoring matrices is the blocks substitution matrix (BLOSUM) series. Henikoff and Henikoff (1992, 1996) used the BLOCKS database, which consisted of over 500 groups of local multiple alignments (blocks) of distantly related protein sequences. Thus the Henikoffs focused on conserved regions (blocks) of proteins that are distantly related to each other.

The BLOCKS database is at ▸http://www.blocks.fhcrc.org. We will describe it further in Chapter 8.

The BLOSUM62 matrix merges all proteins in an alignment that have 62% amino acid identity or greater into one sequence. If a block of aligned RBP orthologs includes several that have 62, 80, and 95% amino acid identity, these would all be grouped as one sequence. Substitution frequencies for the BLOSUM62 matrix are weighted more heavily by protein sequences having less than 62% identity. (Thus, this matrix is useful for scoring proteins that share less than 62% identity.) The BLOSUM62 matrix, which is the default scoring matrix used by most BLAST algorithms (Chapter 4), is shown in Figure 3.17.

The PAM matrix is given as 10 times the log base 10 of the odds ratio. The BLOSUM matrix is given as 2 times the log base 2 of the odds ratio. Thus, BLOSUM scores are not quite as large as they would be if given on the same scale as PAM scores. Practically, this difference in scales is not important because alignment scores are often converted from raw scores to normalized bit scores (Chapter 4).

Henikoff and Henikoff (1992) tested the ability of a series of BLOSUM and PAM matrices to detect proteins in BLAST searches of databases. They found that BLOSUM62 performed slightly better than BLOSUM60 or BLOSUM70 and dramatically better than PAM matrices at identifying various proteins. Their matrices were especially useful for identifying weakly scoring alignments. BLOSUM50 and BLOSUM90 are other commonly used scoring matrices in BLAST searches. (For an alignment of two proteins sharing about 50% identity, try using the BLOSUM50 matrix.)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	4																			
R	-1	5																		
N	-2	0	6																	
D	-2	-2	1	6																
C	0	-3	-3	-3	9															
Q	-1	1	0	0	-3	5														
E	-1	0	0	2	-4	2	5													
G	0	-2	0	-1	-3	-2	-2	6												
H	-2	0	1	-1	-3	0	0	-2	8											
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
K	-1	2	0	-1	-1	1	1	-2	-1	-3	-2	5								
M	-1	-2	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

FIGURE 3.17. *The BLOSUM62 scoring matrix of Henikoff and Henikoff (1992). This matrix merges all proteins in an alignment that have 62% amino acid identity or greater into one sequence. BLOSUM62 performs better than alternative BLOSUM matrices or a variety of PAM matrices at detecting distant relationships between proteins. It is thus the default scoring matrix for most database search programs such as BLAST (Chapter 4). Used with permission.*

The relationships of the PAM and BLOSUM matrices are outlined in Figure 3.18. To summarize, BLOSUM and PAM matrices both use log-odds values in their scoring systems. In each case, when you perform a pairwise sequence alignment (or when you search a query sequence against a database), you specify the exact matrix to use based on the suspected degree of identity between the query and its matches. PAM matrices are based on data from the alignment of closely related protein families, and they involve the assumption that substitution probabilities for highly related proteins (e.g., PAM10) can be extrapolated to probabilities for distantly related proteins (e.g., PAM250). In contrast, the BLOSUM matrices are based on empirical observations of more distantly related protein alignments.

Pairwise Alignment and Limits of Detection

When we compare two protein sequences, how many mutations can occur between them before their differences make them unrecognizable? When we compared human and trout RBPs, it was very easy to see their close relationship (see Fig. 3.8). However, when we compared human RBP4 to bovine β-lactoglobulin, the relationship was much less obvious (Fig. 3.5). Intuitively, at some point two homologous proteins are too divergent to be significantly aligned.

The useful detection limits of pairwise sequence alignment can be explored by comparing the percent identity (and percent divergence) of the two sequences versus their evolutionary distance. Consider two protein sequences, each 100 amino

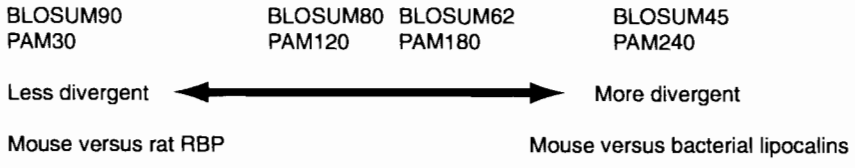


FIGURE 3.18. *Summary of PAM and BLOSUM matrices. High-value BLOSUM matrices and low-value PAM matrices are best suited to study well-conserved proteins such as mouse and rat RBP. BLOSUM matrices with low numbers (e.g., BLOSUM45) or high PAM numbers are best suited to detect distantly related proteins. Remember that in a BLOSUM45 matrix all members of a protein family with greater than 45% amino acid identity are grouped together, allowing the matrix to focus on proteins with less than 45% identity.*

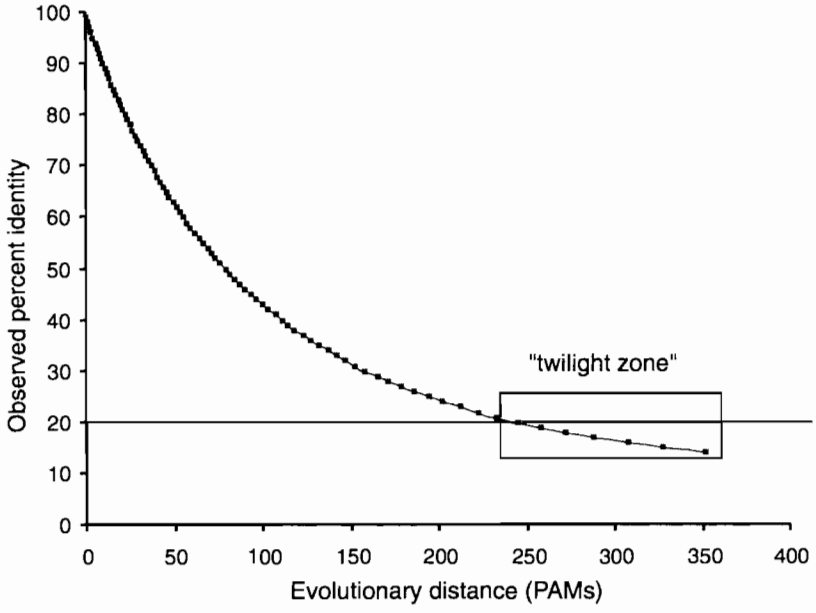


FIGURE 3.19. *Two randomly diverging protein sequences change in a negatively exponential fashion. This plot shows the observed number of amino acid identities per 100 residues of two sequences (y axis) versus the number of changes that must have occurred (the evolutionary distance in PAM units). The twilight zone (Doolittle, 1987) refers to the evolutionary distance corresponding to about 20% identity between two proteins. Proteins with this degree of amino acid sequence identity may be homologous, but such homology is difficult to detect. This figure was constructed using Matlab software with data from Dayhoff (1978) (see Table 3.3).*

A hit is a change in an amino acid residue that occurs by mutation. We will discuss mutations (including multiple hits at a nucleotide position) in Chapter 11 (see Fig. 11.11). We will discuss mutations associated with human disease in Chapter 18.

The plot in Figure 3.19 reaches an asymptote below about 15% amino acid identity. This asymptote would reach about 5% (or the average background frequency of the amino acids) if no gaps were allowed in the comparison between the proteins.

There are about $2^{2n}/\sqrt{\prod n}$ possible global alignments between two sequences of length n (Durbin et al., 2000). (In this equation, $\prod$ is a symbol for product.) For two sequences of length 1000, there are about 10^{600} possible alignments. For two proteins of length 200 amino acid residues (such as lipocalins), the number of possible alignments is over 6×10^{58}.

The number 6×10^{58} represents an exponential (6×10 raised to the fifty-eighth power). This can also be represented as 6E58 or 6exp58.

acids in length, in which one sequence is fixed and various numbers of mutations are introduced into the other sequence. A plot of the two diverging sequences has the form of a negative exponential (Fig. 3.19) (Doolittle, 1987; Dayhoff, 1978). If the two sequences have 100% amino acid identity, they have zero changes per 100 residues. If they share 50% amino acid identity, they have sustained an average of 80 changes per 100 residues. One might have expected 50 changes per 100 residues in the case of two proteins that share 50% amino acid identity. However, any position can be subject to multiple hits. Thus, percent identity is not an exact indicator of the number of mutations that have occurred across a protein sequence. When a protein sustains about 250 hits per 100 amino acids, it may have about 20% identity with the original protein, and it can still be recognizable as significantly related. If a protein sustains 360 changes per 100 residues, it evolves to a point at which the two proteins share about 15% amino acid identity and are no longer recognizable as significantly related.

The PAM250 matrix assumes the occurrence of 250 point mutations per 100 amino acids. As shown in Figure 3.19, this corresponds to the "twilight zone." At this level of divergence, it is usually difficult to assess whether the two proteins are homologous. Other techniques, including multiple sequence alignment (Chapter 10) and structural predictions (Chapter 9), may be useful to assess homology in these cases. PAM matrices are available from PAM1 to PAM250, and a specific number of observed amino acid differences per 100 residues is associated with each PAM matrix (Table 3.3 and Fig. 3.19).

Alignment Algorithms: Global and Local

Our discussion so far has focused on matrices that allow us to score an alignment between two proteins. This involves the generation of scores for identical matches, mismatches, and gaps. We also need an appropriate algorithm to perform the alignment. When two proteins are aligned, there is an enormous number of possible alignments.

There are two main types of alignment: global and local. We will explore these approaches next. A *global alignment* such as the method of Needleman and Wunsch contains the entire sequence of each protein or DNA sequence. We saw a global alignment of RBP4 and β-lactoglobulin in Figure 3.5 above. A *local alignment* such as the method of Smith and Waterman focuses on the regions of greatest similarity between two sequences. For many purposes, a local alignment is preferred, because only a portion of two proteins aligns. (We will study the modular nature of proteins in Chapter 8.) Most database search algorithms, such as BLAST (Chapter 4), use local alignments.

Each of these methods is guaranteed to find one or more optimal solutions to the alignment of two protein or DNA sequences. We will then describe two rapid-search algorithms, BLAST and FASTA. BLAST represents a simplified form of local alignment that is popular because the algorithm is very fast and easily accessible.

Global Sequence Alignment: Algorithm of Needleman and Wunsch

One of the first and most important algorithms for aligning two protein sequences was described by Needleman and Wunsch (1970). This algorithm is important because it produces an optimal alignment of protein or DNA sequences, even allowing the introduction of gaps. The result is optimal, but not all possible alignments need to be evaluated. An exhaustive pairwise comparison would be too computationally expensive to perform.

We can describe the Needleman–Wunsch approach to global sequence alignment in three steps: (1) setting up a matrix, (2) scoring the matrix, and (3) identifying the optimal alignment. We will illustrate this process using the matrix described in the original paper over 30 years ago (Needleman and Wunsch, 1970).

Step 1: Setting Up a Matrix. First, we compare two sequences in a two-dimensional matrix (Fig. 3.20 and following figures). The first sequence, of length m, is listed horizontally along the x axis so that its amino acid residues correspond to the columns. The second sequence, of length n, is listed vertically along the y axis, with its amino acid residues corresponding to rows.

We will describe rules (below) for tracing a diagonal path through this matrix; the path describes the alignment of the two sequences. A perfect alignment between two identical sequences would simply be represented by a diagonal line extending from the top left to the bottom right (Figs. 3.20*a,b*). Any mismatches between two sequences would still be represented on this diagonal path (Fig. 3.20*c*). However, the score that is assigned might be adjusted according to some scoring system. In the example of Fig. 3.20*c*, the mismatch of V and M residues might be assigned a score lower than the perfect match of M and M shown in Figure 3.20*b*.

Gaps are represented in this matrix using horizontal or vertical paths, as shown in Figures 3.20*a,d,e*. Any gap in the top sequence is represented as a vertical line (Figs. 3.20*a,d*), while any gap in the vertical sequence is drawn as a horizontal line (Figs. 3.20*a,e*). These gaps can be of any length. Gaps can be internal or terminal.

We will next set up the matrix needed to compare two sequences as described in the original paper by Needleman and Wunsch (1970) (Fig. 3.21). We will study this example in detail to see how to globally align any two sequences. The first sequence is again listed horizontally along the x axis, while the second sequence is listed vertically along the y axis.

Step 2: Scoring the Matrix. The goal of this algorithm is to identify an optimal alignment. As described next, we will set up two matrices: an amino acid identity

TABLE 3-3 Relationship between Observed Number of Amino Acid Differences per 100 Residues of Two Aligned Protein Sequences and Evolutionary Difference[a]

Observed Differences in 100 Residues	Evolutionary Distance in PAMs
1	1.0
5	5.1
10	10.7
15	16.6
20	23.1
25	30.2
30	38.0
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246

[a]The number of changes that must have occurred, in PAM units. Adapted from Dayhoff (1978, p. 375). Used with permission.

The Needleman and Wunsch approach is an example of a dynamic programming algorithm. It is called "dynamic" because the alignment is created on a residue-by-residue basis in a search for the optimal alignment. The word "programming" refers to the use of a set of rules to determine the alignment.

This algorithm is also sometimes called the Needleman–Wunsch–Sellers algorithm. Sellers (1974) provided a related alignment algorithm (one that focuses on minimizing differences, rather than on maximizing similarities). Smith et al. (1981) showed that the Needleman–Wunsch and Sellers approaches are mathematically equivalent.

The J and B abbreviations are no longer commonly used for amino acids.

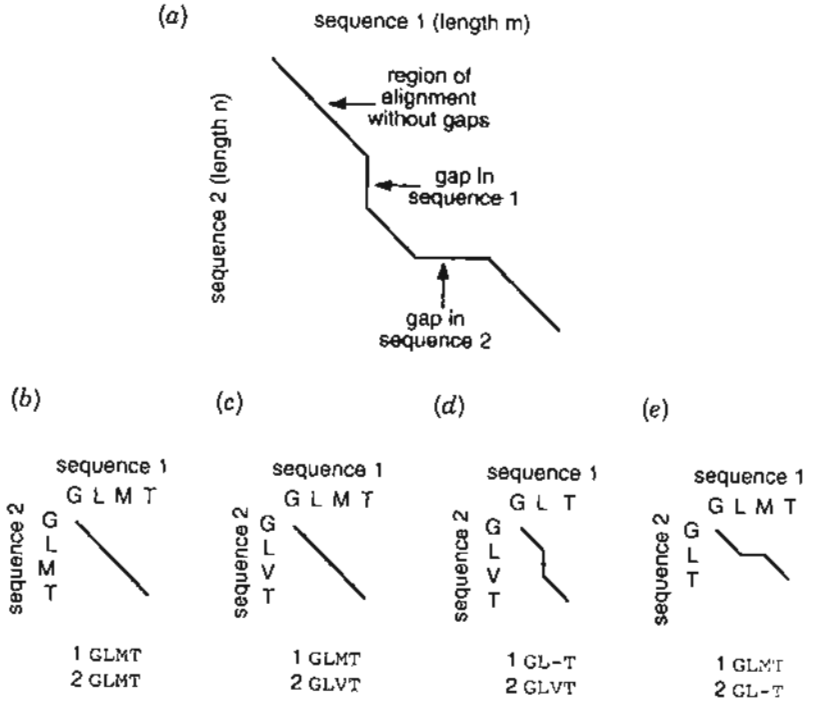


FIGURE 3.20. Pairwise alignment of two amino acid sequences using a dynamic programming algorithm of Needleman and Wunsch (1970) for global alignment. (a) Two sequences can be assigned a diagonal path through the matrix, and when necessary, the path can deviate horizontally or vertically, reflecting gaps that are introduced into the alignment. (b) Two identical sequences form a path on the matrix that fits a diagonal line. (c) If there is a mismatch (or multiple mismatches), the path still follows a diagonal, although a scoring system may penalize the presence of mismatches. If the alignment includes a gap in (d) the first sequence or (e) the second sequence, the path includes a vertical or horizontal line.

matrix and then a scoring matrix. In the scoring matrix, we can employ a simple scoring system in which each amino acid identity gains a score of +1, each mismatch scores −1, and each gap position scores only −2.

In the simple scoring system, insertions and deletions (score −2) are rare compared to amino acid substitutions (score −1), so it makes sense to penalize them more heavily.

We begin by creating an amino acid identity matrix. To do this, we simply place a value of +1 in each cell of the matrix where the two proteins share an identical residue (Fig. 3.21*a*). As described for Figure 3.20*b*, if the two sequences were identical, then the +1 values in this matrix would describe a diagonal line from top left to bottom right.

Note that in linear algebra an identity matrix is a special kind of number matrix that has the number 1 from top left to bottom right. For sequence alignments the amino acid identity matrix is simply a matrix showing all the positions of shared amino acid identity between two sequences, as shown in Figure 3.21*a*.

Next, we set up a scoring matrix (Figs. 3.21*b–d*). Our goal in finding an optimal alignment is to determine the path through the matrix that maximizes the score. This usually entails finding a path through as many positions of identity as possible while introducing as few gaps as possible. There are four possible occurrences at each position i,j (i.e., in each cell in the matrix):

1. Two residues may be perfectly matched (i.e., identical).
2. They may be mismatched.
3. A gap may be introduced from the first sequence.
4. A gap may be introduced from the second sequence.

The Needleman and Wunsch algorithm provides a score, corresponding to each of these possible outcomes, for each position of the aligned sequences. The algorithm also specifies a set of rules describing how we can move through the matrix.

The assignment of scores begins at the bottom right of the matrix, corresponding to the carboxy termini of the proteins. Each row is filled in with values, and we fill in the rows, one cell at a time, beginning with the bottom row at the right and proceeding to the top. Along the bottom row, we can fill in zeros, leaving a value of +1 in the cell containing the P × P identity (Fig. 3.21*b*). We know that a +1 belongs in the P × P cell because we can refer to the amino acid identity matrix (Fig. 3.21*a*).

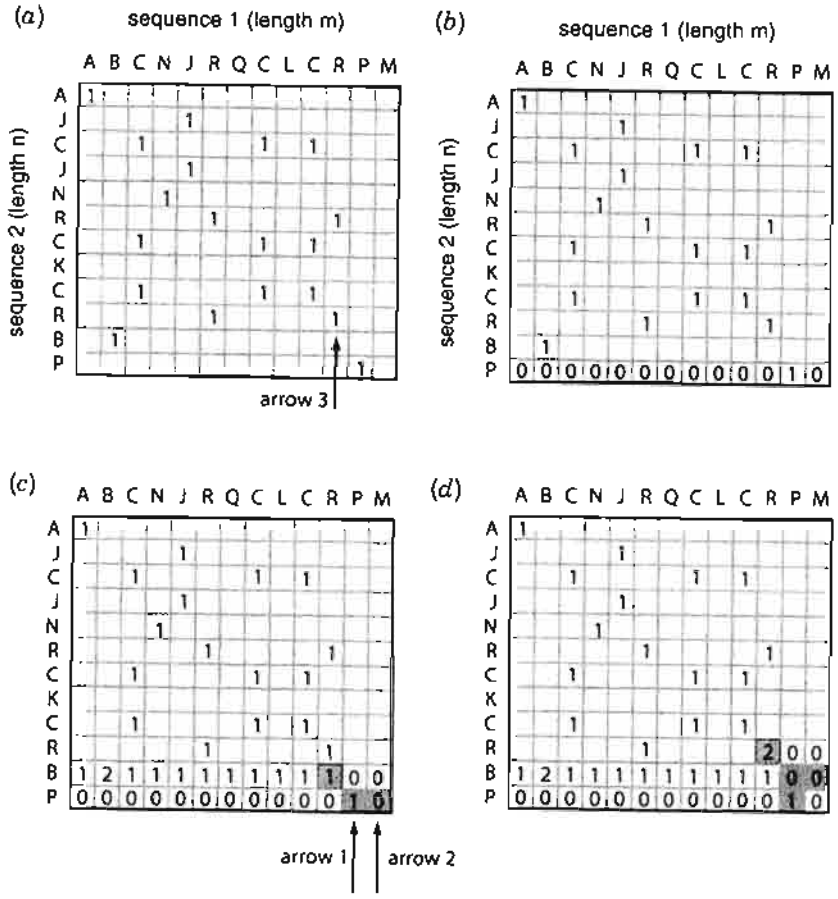


FIGURE 3.21. Pairwise alignment of two amino acid sequences using a dynamic programming algorithm of Needleman and Wunsch (1970) for global alignment. The sequences in this example are from the original paper. (a) We first assign a score of +1 for each position of identity in the amino acid identity matrix. All other cells are assigned a value of zero (not shown). We go to the carboxy terminus of each protein (lower right corner of the matrix) and compute the matrix row by row. (b) Along the bottom row, we can fill in the zeros, leaving a +1 in the P × P cell. The rule to follow is that each cell receives its +1 score [if any, from (a)] plus the maximum value of the cell diagonally below to the right or along that corresponding row and column. In this figure and Figure 3.22, this means that we select the maximum value from the cells shaded gray. (c) On the second row from the bottom, the cell B × R (shaded red) receives a score of +1 because that value is the maximal score that is found diagonally (in the P × P cell, shaded gray) or in the corresponding row (gray box). That value of +1 is also added to all the cells along the second row to the left of the B × R position. Note that in (c) the position of identity where B matches B now receives a score of 2 (+1 from its identity and +1 from the P × P match of the bottom row). In (d) consider the R × R position (shaded red). Its score is +1 (from the identity matrix) plus the maximum of the three scores (diagonally below to the right, shaded gray; add +1). Used with permission.

How can we move through the matrix, defining a path that will correspond to the sequence alignment? Let us define the matrix as having i corresponding to rows and j corresponding to columns. A cell can be defined as having a position (i, j). Consider any cell for which we want to find the score; as an example, look at the cell that is shaded red in Figure 3.21*c*. There are several rules for movement:

- First, both i and j must increase. This means that to calculate the score in the red box, we must look down and to the right in the matrix. This is reasonable; it would not make sense to be able to violate the linear arrangement of amino acids (or nucleotides) in a sequence.

- Either i or j must increase by 1, but the other can increase by more than 1. In this example, i can only increase by 1. The cell for j can increase by 1 (Fig. 3.21*c*, gray cell at arrow 1) or it can increase by 2 (gray cell at arrow 2).

We score each cell in two steps. We inspect the amino acid identity matrix (Fig. 3.21*c*) and take its value [in the case of the red-shaded cell in (*c*), that value is zero]. We then add the maximum value we can obtain using the two movement rules. In this case, that value is 1 (Fig. 3.21*c*, arrow 1). Thus, we place the value 1 in the red-shaded cell of Figure 3.21*c*.

It is even easier to see these movement rules if we examine other cells as the matrix is filled in. For the cell shaded red in Figure 3.21*d*, the score is +1 from the amino acid identity matrix (Fig. 3.21*a*, arrow 3) plus the maximum value from our movement rules, which is +1. Thus we place the value 2 in that red-shaded cell. The next cell (Fig. 3.22*a*) receives a score of zero from the amino acid identity

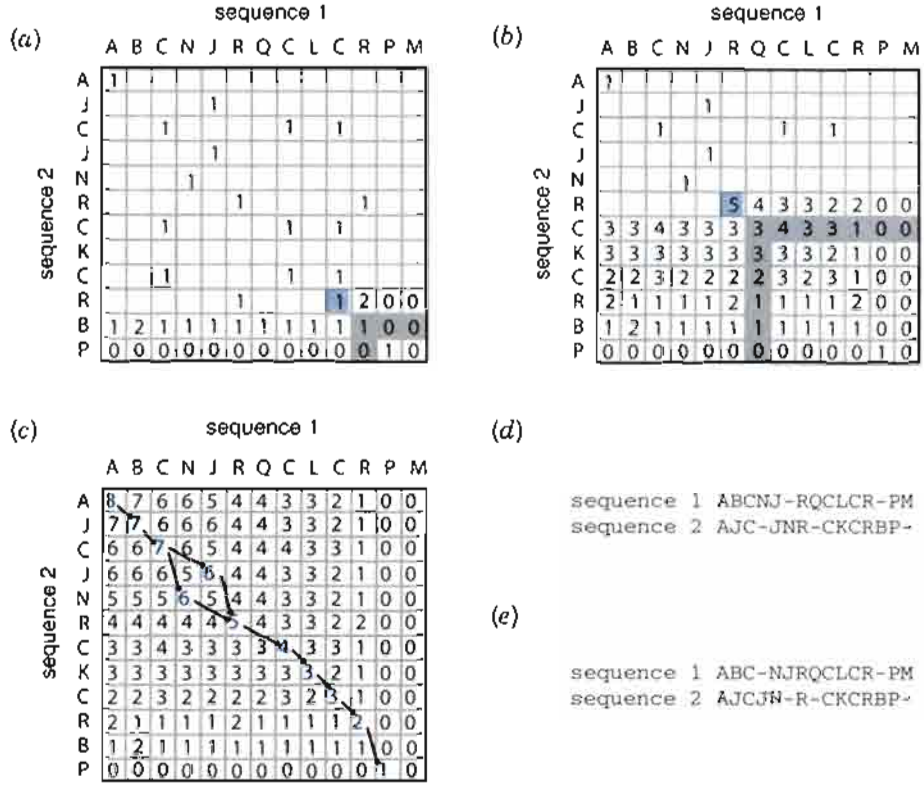


FIGURE 3.22. Global pairwise alignment of two amino acid sequences using a dynamic programming algorithm: scoring the matrix and using the trace-back procedure to obtain the alignments. The procedure for scoring the matrix (described in Fig. 3.21) is continued up each row. (a) At the red-shaded cell, the score is 0 (there is no +1 identity score) plus the maximum in the gray cells (+1). (b) At the R × R cell (shaded red), the score is +1 [for the identity, as seen in (a)] plus the maximum of three scores: first, at $(i + 1, j + 1)$ (i.e., score = +3); second, $i + 2$, $j + 1$ to the end of that row i (score = +4); or third, $i + 1$, $j + 2$ to the end of that column. Thus the optimal score is $+1 + 4 = 5$. (c) In this way, the matrix is filled. The trace-back procedure is then used to define the optimal alignment path(s). To do this, begin at the amino terminus of each protein (where the score is +8) and follow the scores down through the matrix. Here the first three residues are aligned without a gap (ABC in sequence 1 matches AJC in sequence 2). At that point there are two equally optimal paths, each with a score of +6, each of which introduces a gap in one or the other sequence. The upper path is shown in (d), and the lower path is shown in (e). Since this is a global alignment, the complete sequences 1 and 2 are aligned, rather than just an internal segment. Used with permission.

matrix plus the maximum of the cells shaded gray (i.e., one), for a total score of 1.

The example of the red-shaded cell in Figure 3.22*b* yields a score of +5. This consists of +1 (from the amino acid identity matrix) and +4 (from the maximum value available according to our rules of movement). This +4 value is at position $i + 1$, $j + 2$. The physical interpretation of this "jump" (which scores higher than the +3 available from the diagonal choice $i + 1$, $j + 1$) is that the introduction of a gap in an alignment of sequence 1 (residues RQC in Figs. 3.22*c*, *d*, *e*) and sequence 2 (residues R–C) will yield a high score and provide an optimal alignment.

In summary, at each position i, j take the value in that cell plus the maximum score obtained from any one of these three values:

1. Identify the score diagonally down (at position $i + 1$, $j + 1$), without including any gaps.
2. Find the highest score in position $i + 1$, $j + 2$ to the end of row j. This corresponds to the addition of a gap in the column. This gap could be one character or it could be larger.
3. Find the highest score in position $i + 2$, j moving to the end of column i. This corresponds to the addition of a gap in the row.

Step 3: Identifying the Optimal Alignment. After the matrix is filled, the alignment is determined by a trace-back procedure. For this, we begin at the upper left of the matrix (amino termini of the proteins). Begin with the highest value, +8 (Fig. 3.22*c*). This corresponds to an alignment of residues A to A in the two sequences (Figs. 3.22*a–e*). We find the path down and to the right with the highest numbers. The next alignment is a mismatch between J and B (with a score of 7); next is an identity between two C residues. In each case we are searching for the highest score that is available one cell down along the diagonal. In the next position, it would be possible to align J to N along the diagonal, but that only corresponds to a score of +5; a higher score (+6) is achieved by choosing a cell directly to the right (see Fig. 3.22*d*) or directly down (Fig. 3.22*e*). Going off the diagonal automatically implies the insertion of a gap in one of the sequences. This results in some penalty. There may be more than one optimal alignment, as shown in this example where the two solutions (Figs. 3.22*d*, *e*) have an equally high score. Here the scoring system consists simply of +1 for a match. However, the introduction of a sophisticated scoring matrix (such as BLOSUM62 or PAM250) makes it unlikely that there will be multiple optimal alignments.

A variety of programs implement global alignment algorithms (see Web Resources at the end of this chapter). An example is the GAP program of GCG (Box 3.6). Two sequences are entered, and penalties are selected for gap creation and extension (Fig. 3.23). The resulting global alignment includes descriptions of the percent identity and similarity shared by the two proteins, the length of the alignment, and the number of gaps introduced (Fig. 3.24).

The Needleman–Wunsch algorithm is an example of dynamic programming (Sedgewick, 1988). This means that an optimal path (i.e., an optimal alignment) is detected by incrementally extending optimal subpaths—that is, by making a series of decisions at each step of the alignment as to which pair of residues corresponds to the best score. The overall goal is to find the path moving along the diagonal of the matrix that lets us obtain the maximal score. This path specifies the optimal alignment.

BOX 3-6
Genetics Computer Group

The Genetics Computer Group (GCG) is also known as the "Wisconsin package" (▸http://www.accelrys.com) (see Appendix). GCG is a collection of 130 algorithms for the analysis of DNA, RNA, and protein sequences. There are presently over 40,000 users of GCG worldwide. This is a commercial package that is available in Unix and web-based versions.

To use GCG, one first deposits sequences into the main sequence editor, "seqed." Protein and DNA sequences can be entered manually or can be obtained with the "fetch" command. (For all GCG commands, extensive help menus are available through the "genhelp" tool.)

Once the relevant sequences are available in a GCG directory, pairwise alignment can be performed using the "GAP" tool (for global sequence alignment using the Needleman and Wunsch algorithm) or the "Bestfit" tool (for local sequence alignment using the Smith and Waterman algorithm). We first perform the global alignment with GAP (see Fig. 3.29 below). Note that the program allows the users to specify the range of amino acids to be compared as well as penalties for gap weight and gap length. The output (see Fig. 3.30 later) includes a quality score, which is based upon the scoring matrix that is used; a higher quality score corresponds to a more significant match. Other data include the percent amino acid identity, the number of gaps, and the alignment itself. For a global alignment, the program is constrained to show the entire sequence of both proteins, unless (as in this case) there is no match at the carboxy terminus of a protein.

Local pairwise alignment is performed in a similar fashion, using the "Bestfit" program. We can enter the command line "bestfit hsrbp.pep btlacto.pep -ran=100," which instructs GCG to run the bestfit program, comparing our two proteins. The command "ran=100" specifies that the second sequence should be shuffled randomly and compared 100 times to the first sequence. We note several features about the output (see Fig. 3.26 later). The alignment is now local, so the alignment shows only the most well-matched portions of the pairwise alignment. The aligned region is shorter than for a global alignment, but the percent identity of the matched region is greater. Global and local alignment algorithms often generate dramatically different results. This is true for both pairwise alignments and multiple sequence alignments (Chapter 10).

```
> gap

Gap uses the algorithm of Needleman and Wunsch to find the alignment of
two complete sequences that maximizes the number of matches and minimizes
the number of gaps

 GAP of what sequence 1 ?  hsrbp.pep

                 Begin (* 1 *) ?
                End (*   199 *) ?

 to what sequence 2 (* hsrbp.pep *) ?  btlacto.pep

                 Begin (* 1 *) ?
                End (*   178 *) ?

 What is the gap creation penalty (* 8 *) ?

 What is the gap extension penalty (* 2 *) ?

 What should I call the paired output display file (* hsrbp.pair *) ?

 Aligning ........-.
 Aligning ........-.

         Gaps:     0
      Quality:    37
 Quality Ratio: 0.200
  % Similarity: 31.902
       Length:   214
```

FIGURE 3.23. Global pairwise alignment is performed with the GAP program of GCG. In this example, we compare human RBP4 (hsrbp.pep) to bovine β-lactoglobulin (btlacto.pep). These sequences were copied into the GCG sequence editor ("seqed") from the NCBI Entrez protein site and named with suffixes (.pep) to indicate that they are proteins rather than nucleic acids. Once entered in GCG, the sequences can be analyzed using several dozen different algorithms, including GAP. Here, the regions of the proteins to be compared can be selected as well as gap creation and gap extension penalties.

FIGURE 3.24. Output of the GAP program shows a global alignment of two proteins, RBP4 and β-lactoglobulin. The output includes a description of the score (quality = 37), the number of gaps, and the percent identity. In the pairwise alignment, the first sequence (RBP4) is depicted on top.

Local Sequence Alignment: Smith and Waterman Algorithm

The local alignment algorithm of Smith and Waterman (1981) is the most rigorous method by which subsets of two protein or DNA sequences can be aligned. Local alignment is extremely useful in a variety of applications such as database searching where we may wish to align domains of proteins (but not the entire sequences). A local sequence alignment algorithm resembles that for global alignment in that two proteins are arranged in a matrix and an optimal path along a diagonal is sought. However, there is no penalty for starting the alignment at some internal position, and the alignment does not necessarily extend to the ends of the two sequences.

For the Smith-Waterman algorithm a matrix is constructed with an extra row along the top and an extra column on the left side. Thus for sequences of lengths m and n, the matrix has dimensions $m + 1$ and $n + 1$. The rules for defining the value in each position of the matrix differ slightly from those used in the Needleman-Wunsch algorithm. The score in each cell is selected as the maximum of the preceding diagonal or the score obtained from the introduction of a gap. However, the score cannot be negative: A rule introduced by the Smith-Waterman algorithm is that if all other score options produce a negative value, then a zero must be inserted in the cell. Thus the score $S(i,j)$ is given as the maximum of four possible values (Fig. 3.25):

1. The score from the cell at position $i - 1$, $j - 1$; that is, the score diagonally up to the left. To this score, add the new score at position $s[i, j]$, which consists of either a match or a mismatch.
2. $s(i, j - 1)$ (i.e., the score one cell to the left) minus a gap penalty.
3. $s(i - 1, j)$ (i.e., the score immediately above the new cell) minus a gap penalty.
4. The number zero. This condition assures that there are no negative values in the matrix. In contrast negative numbers commonly occur in global alignments because of gap or mismatch penalties (note the log-odds matrices in this chapter).

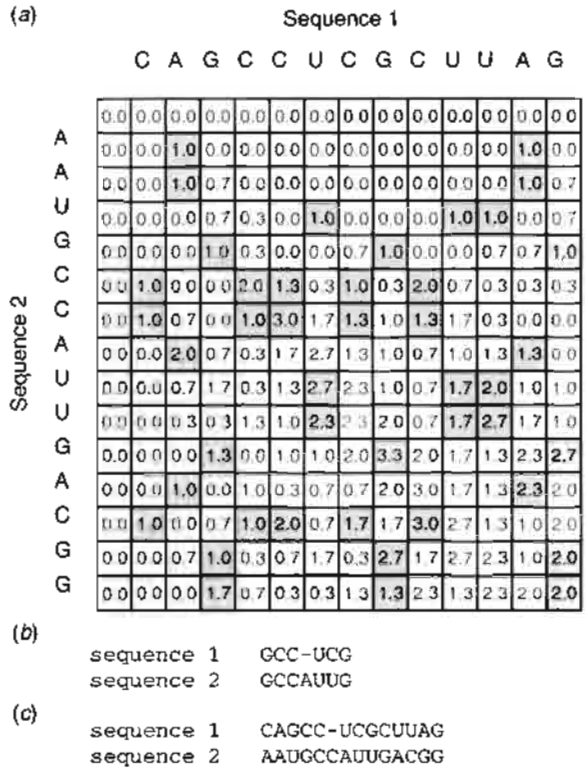


FIGURE 3.25. Local sequence alignment method of Smith and Waterman (1981). (a) In this example, the matrix is formed from two RNA sequences (CAGCCUCGCUUAG and AAUGCCAUUGACGG). While this is not an identity matrix (such as that shown in Figure 3.21a), positions of nucleotide identity are shaded gray. They scoring system here is +1 for a match, minus one-third for a mismatch, and a gap penalty of the difference between a match and a mismatch (−1.3 for a gap of length one). The matrix is scored according to the rules outlined on the bottom of page 69. The highest value in the matrix (3.3) corresponds to the beginning of the optimal local alignment, and the aligned residues (shaded red) extend up and to the left until a value of zero is reached. (b) The local alignment derived from this matrix is shown. Note that this alignment includes identities, a mismatch, and a gap. (c) A global alignment of the two sequences is shown for comparison to the local alignment. Note that it encompasses the entirety of both sequences. Used with permission.

An example of the use of a local alignment algorithm to align two nucleic acid sequences adapted from Smith and Waterman (1981), is shown in Figure 3.25. The topmost row and the leftmost column of the matrix are filled with zeros. The maximal alignment can begin and end anywhere in the matrix (within reason; the linear order of the two amino acid sequences cannot be violated). The procedure is to identify the highest value in the matrix (this value is 3.3 in Figure 3.25a). This represents the end (carboxy-terminal portion) of the alignment. This position is not necessarily at the lower right corner, as it must be for a global alignment. The trace-back procedure begins with this highest value position and proceeds diagonally up to the left until a cell is reached with a value of zero. This defines the start of the alignment, and it is not necessarily at the extreme top left of the matrix.

A requirement of the Smith–Waterman algorithm is that the expected score for a random match is negative. This condition ensures that alignments between very long unrelated sequences do not accrue high scores. Such sequences could otherwise produce spurious alignments having higher scores than the authentic match between two proteins over a shorter region.

An example of a local alignment using the Smith–Waterman algorithm is shown for RBP4 and β-lactoglobulin in Figure 3.26. Compare this with the global

FIGURE 3.26. *Local aligment with the Bestfit program of GCG shows limited portions of a pairwise alignment. Compare this alignment to Figure 3.24, where the global alignment algorithm GAP constrained the output to contain the entire sequence of both proteins. Here, the length of the alignment is shorter, but the percent identity is higher. Note that the "–ran = 100" command was used to generate a quality score (plus or minus the standard deviation) for 100 random shufflings of the β-lactoglobulin sequence. See Figure 3.30.*

alignment of Figure 3.24 and note that the aligned region is shorter for the local alignment, while the percent identity and similarity are higher.

This local alignment is performed using the Bestfit program of GCG (▸http://www.accelrys.com).

Rapid, Heuristic Versions of Smith–Waterman: FASTA and BLAST

While the Smith–Waterman algorithm is guaranteed to find the optimal alignment(s) between two sequences, it suffers from the fact that it is relatively slow. For pairwise alignment, speed is usually not a problem. But when a pairwise alignment algorithm is applied to the problem of comparing one sequence (a "query") to an entire database, the speed of the algorithm becomes a significant issue and may vary by orders of magnitude.

Most algorithms have a parameter N that refers to the number of data items to be processed (see Sedgewick, 1988). This parameter can greatly affect the time required for the algorithm to perform a task. If the running time is proportional to N, then doubling N doubles the running time. If the running time is quadratic (N^2), then for $N = 1000$, the running time is one million. For both the Needleman-Wunsch and the Smith-Waterman algorithms, both the computer space and the time required to align two sequences is proportional to at least the length of the two query sequences multiplied against each other, $m \times n$. For the search of a database of size N, this is $m \times N$.

The modified alignment algorithms introduced by Gotoh (1982) and Myers and Miller (1988) require only $O(nm)$ time and occupy $O(n)$ in space. Instead of committing the entire matrix to memory, the algorithms ignore scores below a threshold in order to focus on the maximum scores that are achieved during the search.

Another useful descriptor is *O*-notation (called "big-Oh notation") which allows one to approximate the upper bounds on the running time of an algorithm. The Needleman-Wunsch algorithm requires $O(mn)$ steps, while the Smith-Waterman algorithm requires $O(m^2n)$ steps. Subsequently, Gotoh (1982) and Myers and Miller (1988) improved the algorithms so they require less time and space.

Two popular local alignment algorithms have been developed that provide rapid alternatives to Smith–Waterman: FASTA (Pearson and Lipman, 1988) and BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990). Each of these algorithms requires less time to perform an alignment. The time saving occurs

FASTA stands for FAST-All, referring to its ability to perform a fast alignment of all sequences (i.e. proteins or nucleotides).

because FASTA and BLAST restrict the search by scanning a database for likely matches before performing more rigorous alignments. These are heuristic algorithms (Box 3.2) that sacrifice some sensitivity in exchange for speed; in contrast to Smith-Waterman, they are not guaranteed to find optimal alignments.

The FASTA search algorithm introduced by Pearson and Lipman (1988) proceeds in four steps.

The parameter *ktup* refers to multiples such as duplicate, triplicate, or quadruplicate (for $k = 2, k = 3, k = 4$). The *ktup* values are usually 3–6 for nucleotide sequences and 1–2 for amino acid sequences. A small *ktup* value yields a more sensitive search but requires more time to complete.

1. A lookup table is generated consisting of short stretches of amino acids or nucleotides from a database. The size of these stretches is determined from the *ktup* parameter. If *ktup* = 3 for a protein search, then the query sequence is examined in blocks of three amino acids against matches of three amino acids found in the lookup table. The FASTA program identifies the 10 highest scoring segments that align for a given *ktup*.
2. These 10 aligned regions are rescored, allowing for conservative replacements, using a scoring matrix such as PAM250.
3. High-scoring regions are joined together if they are part of the same proteins.
4. FASTA then performs a global (Needleman-Wunsch) or local (Smith-Waterman) alignment on the highest scoring sequences, thus optimizing the alignments of the query sequence with the best database matches. Thus dynamic programming is applied to the database search in a limited fashion, allowing FASTA to return its results very rapidly because it evaluates only a portion of the potential alignments.

William Pearson of the University of Virginia provides FASTA online. Visit ▸http://alpha10.bioch.virginia.edu/fasta/. Another place to try FASTA is at the European Bioinformatics Institute website, ▸http://www.ebi.ac.uk/fasta33/.

Basic Local Alignment Search Tool (BLAST)

BLAST was introduced as a local alignment search tool that identifies alignments between a query sequence and a database without the introduction of gaps (Altschul et al., 1990). The version of BLAST that is available today allows gaps in the alignment. We will introduce BLAST in more detail in Chapter 4, where we will describe the heuristic algorithm. Here we will introduce the practical use of "BLAST 2 Sequences," which is the part of the BLAST website at NCBI that permits local pairwise alignment of two sequences (Tatusova and Madden, 1999). It is easy to perform pairwise alignments of proteins (or DNA sequences) using this program. Perform the following steps:

From the main NCBI web page, click BLAST from the top bar (or go directly to ▸http://www.ncbi.nlm.nih.gov/BLAST/). Select "BLAST 2 Sequences."

1. Choose the program blastn (for "BLAST nucleotides") if you want to compare two DNA sequences or blastp (for "BLAST proteins") if you have proteins. We will use proteins in our examples.
2. Enter the sequences or their accession numbers. For example, enter the accession numbers of human RBP (NP_006735) and β-lactoglobulin (P02754) (Fig. 3.27).
3. Select any optional parameters.
   - You can choose from six scoring matrices: BLOSUM62, BLOSUM50, BLOSUM90, PAM250, PAM70, PAM30. Select PAM250.
   - You can change the gap creation penalty and gap extension penalty.
   - For blastn searches you can change reward and penalty values.
   - There are other parameters you can change, such as word size, expect value, filtering, and dropoff values. We will discuss these more in Chapter 4.

FIGURE 3.27. *Example of pairwise sequence alignment using the NCBI BLAST 2 Sequences algorithm. The accession numbers of human RBP4 (NP_006735) and bovine β-lactoglobulin (P02754) are entered. It is important to (1) choose the appropriate program (blastp for protein searches, blastn for DNA searches), (2) enter the sequences in FASTA form or (as in this case) using the accession numbers, (3) select optional parameters such as gap penalties, and (4) click "Align."*

4. Click "align." A typical result is shown in Figure 3.28. Note that for distantly related proteins such as these, an alignment is only returned if the appropriate scoring matrix (PAM250) and gap penalties are assigned. The output includes a dot plot in which aligned portions of the two sequences are represented along a diagonal (Box 3.7). The output also includes a pairwise alignment using the single-letter amino acid code such as we have seen in Figure 3.5.

SIGNIFICANCE OF PAIRWISE ALIGNMENTS: PERCENT IDENTITY

How can we decide whether two sequences are significantly related from an evolutionary point of view? Could the alignment of RBP with β-lactoglobulin, which yielded 26% amino acid identity, have occurred by chance?

A rule of thumb is that if two proteins share 25% or more amino acid identity over a span of 150 or more amino acids, they are probably significantly related (Brenner et al., 1998). If we consider an alignment of just 70 amino acids, it is popular to consider the two sequences "significantly related" if they share 25%

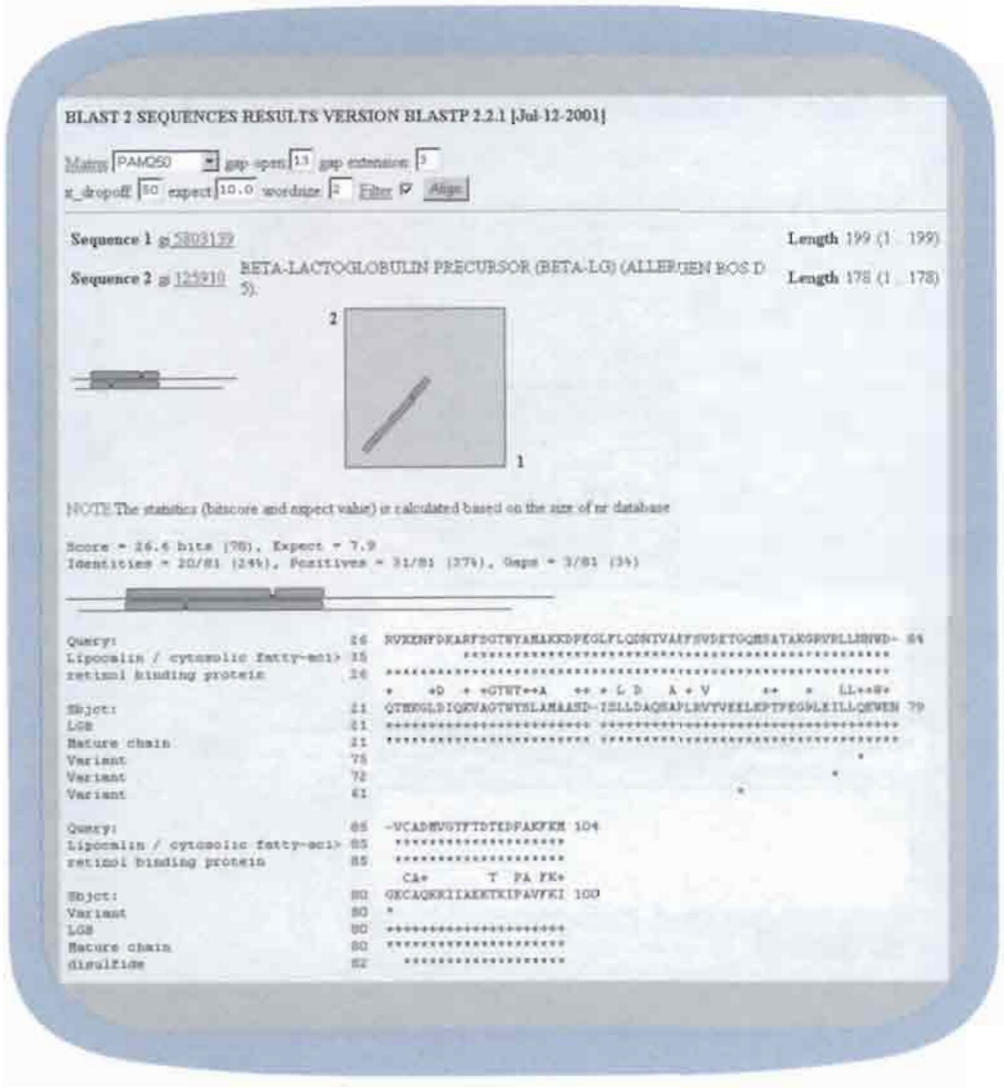


FIGURE 3.28. Example of an output of the BLAST 2 Sequences program. Note that the PAM matrix is set to 250 for these distantly related proteins, and the gap open and gap extension penalties have been adjusted to 13 and 3. No significant alignment is detected using PAM30, PAM70, or BLOSUM matrices. The output includes a dot plot showing an overview of the pairwise alignment as well as an alignment of the amino acid sequences. Note that the output includes bit scores, an expect value, and descriptions of identities, positives, and gaps. We will describe these outputs in Chapter 4.

amino acid identity. However, Brenner et al. (1998) have shown that this may be erroneous, in part because the enormous size of today's molecular sequence databases increases the likelihood that such alignments occur by chance. For an alignment of 70 amino acid residues, 40% amino acid identity is a reasonable threshold to estimate that two proteins are homologous (Brenner et al., 1998). If two proteins share about 20–25% identity over a reasonably long stretch (e.g., 70–100 amino acid residues), they are in the "twilight zone" (Fig. 3.19), and it is more difficult to be sure. Two proteins that are completely unrelated often share about 10–20% identity when aligned. This is especially true because the insertion of gaps can greatly improve the alignment of any two sequences.

We will see in Chapter 10 that multiple sequence alignments can offer three times more sensitivity than pairwise sequence alignment. We will also see in

BOX 3-7
Dot Plots

The dot plot or dot matrix is a graphical method for comparing two sequences. One protein or nucleic acid sequence is placed along the x axis, and the other is placed along the y axis. Positions of identity are scored with a dot. A region of identity between two sequences results in the formation of a diagonal line. There may be several diagonals. Also, the parameters may be adjusted to show a dot only when a group of residues (e.g., five nucleotides in a row) are matched between the two sequences. If a protein (or nucleic acid) sequence is compared to itself with a dot plot, repetitive domains within the sequence may be easily visualized.

We will encounter dot plots in Chapter 13 when we compare bacterial genome sequences to each other. We will also see a dot plot in Chapter 15 (on fungi). Two *Saccharomyces cerevisiae* chromosome sequences were BLAST searched against each other. The resulting dot plot showed many diagonal lines indicating homologous regions. This led researchers to propose that the entire yeast genome duplicated over 100 MYA.

Dotlet is a web-based diagonal plot tool available from the Swiss Institute of Bioinformatics (▸http:www.isrec.isb-sib.ch/java/dotlet/Dotlet.html).

Chapter 5 that scoring matrices ("profiles") can be customized to a sequence alignment, greatly increasing the sensitivity of a search.

Tests for Statistical Significance of Pairwise Alignments

If two proteins share limited amino acid identity (e.g., 20–25%), how can we determine whether they are significantly related? Alignment algorithms report the score of a pairwise alignment or the score of the best alignments of a query sequence against an entire database of sequences (Chapter 4). We need statistical tests to decide whether the matches are true positives (i.e., whether the two aligned proteins are genuinely homologous) or whether they are false positives (i.e., whether they have been aligned by the algorithm by chance) (Fig. 3.29). For the alignments that are not reported by an algorithm, for instance because they fall below some scoring threshold, we would like to evaluate whether the sequences are true negatives (i.e., genuinely unrelated) or whether they are false negatives, that is, homologous sequences that receive a score suggesting that they are not homologous.

A main goal of alignment algorithms is thus to maximize the sensitivity and specificity of sequence alignments (Fig. 3.29). Sensitivity is the number of true positives divided by the sum of true-positive and false-negative results. This is a measure of the ability of an algorithm to correctly identify genuinely related sequences. Specificity is the number of true-negative results divided by the sum of true-negative and false-positive results. This describes the sequence alignments that are not homologous.

There are several principal approaches to deciding whether an alignment is statistically significant. A protein such as RBP4 can be compared to another, such as β-lactoglobulin, and we can generate a raw score using some scoring system. To assess the significance of the alignment, we can compare the result to the outcome of comparing RBP to other sequences, such as:

- Many other proteins that are known to not be homologous to RBP4
- Other sequences, such as β-lactoglobulin itself, that have been scrambled to preserve their compositional properties (length, amino acid composition)
- Randomly generated sequences

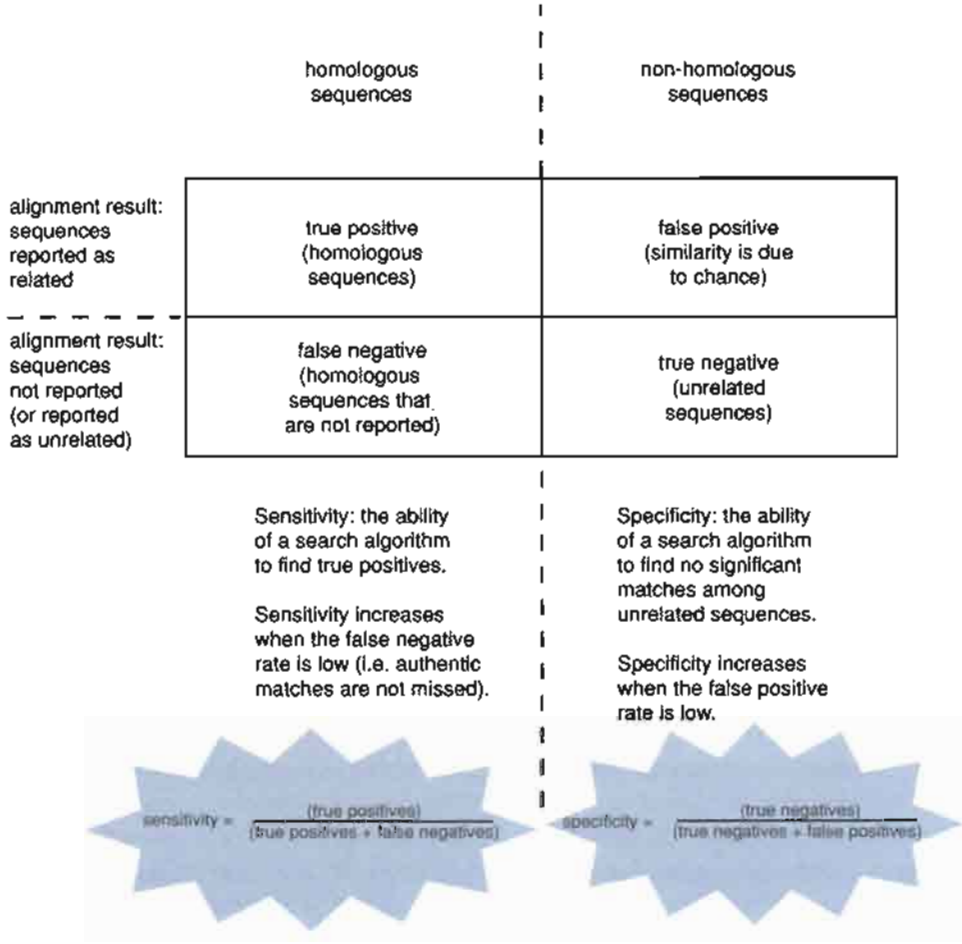


FIGURE 3.29. *Sequences alignments, whether pairwise (this chapter) or from a database search (Chapter 4), can be classified as true or false and positives or negatives. Statistical analyses of alignments provide the main way that you can evaluate whether an alignment represents a true positive, that is, an alignment of homologous sequences. Ideally, an alignment algorithm can maximize both sensitivity and specificity.*

We will approach this problem by considering global then local alignments. In Chapter 4, we will extend these statistical arguments to the case of database searches using tools such as BLAST.

Statistical Significance of Global Alignments

The statistical properties of global alignments are poorly understood (relative to local alignment statistics), but nonetheless we can consider how to assess their significance. Let us continue our example of RBP4 aligned to β-lactoglobulin. We can first align them and obtain a score. We can then scramble (or shuffle) the β-lactoglobulin sequence 100 times, perform 100 alignments, record the scores, and then ask how the authentic score compares to the mean randomized score. We can express the authentic score in terms of how many standard deviations above the mean it is. A Z score is calculated as:

$$Z = \frac{x - \mu}{\sigma} \tag{3.3}$$

where x is the current score of two aligned sequences, μ is the mean score of many sequence comparisons using a scrambled sequence, and σ is the standard deviation of those measurements obtained with random sequences.

What can we conclude from this Z score? If 100 alignments of shuffled proteins all have a score less than the authentic score of RBP4 and β-lactoglobulin, this indicates that the probability (p) value is less than 0.01 that this occurred by chance.

Consider the case of comparing RBP to 1 million proteins in a database, and not just to β-lactoglobulin. In that situation, it is not appropriate to set the α level (i.e., the probability at which you reject the null hypothesis) to the traditional $p < 0.05$. Instead, α should be divided by the number of trials, and in this example you would define probabilities less than $0.05/10^6$, or 5×10^{-8}, as significant. This is called a Bonferroni correction (see Chapter 7). It is possible to have a large Z score from a pairwise alignment (e.g., 10, indicating that this score is 10 standard deviations above the mean) without the alignment having statistical significance.

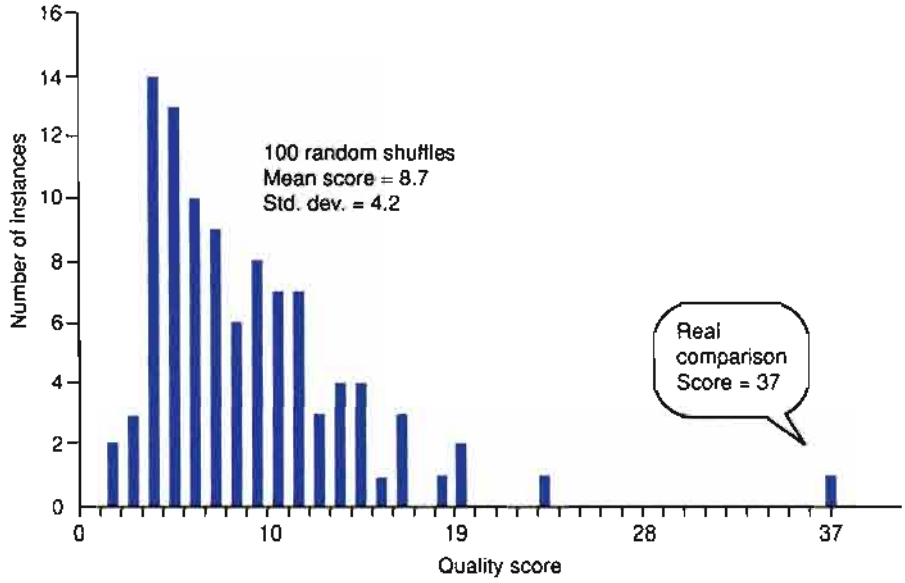


FIGURE 3.30. *Evaluation of the statistical significance of a pairwise sequence alignment using a randomization test. Human RBP4 was aligned to bovine β-lactoglobulin using the GAP program of GCG. The alignment yielded a score of 37. Then the sequence of β-lactoglobulin was shuffled randomly 100 times, maintaining the length and amino acid composition of the protein. Each of the scores was plotted; the mean and standard deviation of the shuffled sequences' scores were 8.7 ± 4.2. Assuming that the randomized scores are normally distributed, it is possible to calculate a Z score to assess the likelihood that the real comparison score occurred by chance.*

(Thus we can reject the null hypothesis that RBP4 and β-lactoglobulin sequences are not significantly related.)

We can do the shuffle test using a computer algorithm such as GAP (from the GCG software package). GAP calculates the score of a global pairwise alignment, which in this case is 37 (Fig. 3.24). The program also performs as many comparisons of RBP to a randomized (jumbled) version of β-lactoglobulin as you would like. In this case, the mean score from 100 randomizations of β-lactoglobulin is 8.7 ± 4.2 (Fig. 3.30). We can calculate a Z score as $(37 - 8.7)/4.2 = 6.7$.

We would like to relate a Z score to a probability value to assess statistical significance. However, it is not possible to do this accurately unless we understand the overall distribution of randomized scores. For global alignments, this distribution is not well characterized, and thus a Z score cannot be converted into a probability value. We will see that for local alignments the statistical theory is far better developed.

Statistical Significance of Local Alignments

Most database search programs such as BLAST (Chapter 4) depend on local alignments. Additionally, many pairwise alignment programs compare two sequences using local alignment. The locally aligned regions are called high-scoring segment pairs, or HSPs. The most important statistic associated with BLAST is the expect (E) value, which describes the number of hits one can expect to obtain by chance when searching a database of a particular size. The E value is a measure of the statistical significance of an alignment result. We will explore it in more detail in Chapter 4.

PERSPECTIVE

The pairwise alignment of DNA or protein sequences is one of the most fundamental operations of bioinformatics. Pairwise alignment allows one to determine the relationship between any two sequences, and the degree of relatedness that is

observed helps one to form a hypothesis about whether they are homologous (descended from a common evolutionary ancestor). Almost all of the topics in the rest of this book are heavily dependent upon sequence alignment. In Chapter 4, we will introduce the searching of large DNA and/or protein databases with a query sequence. Database searching typically involves an extremely large series of local pairwise alignments, with results returned as a rank order beginning with most related sequences.

The algorithms used to perform pairwise alignment were developed in the 1970s, beginning with the global alignment procedure of Needleman and Wunsch (1970). Dayhoff (1978) introduced PAM scoring matrices that permit the comparison and evaluation of distantly related molecular sequences. By the 1980s, local alignment algorithms were introduced (see especially the work of Sellers (1974), Smith and Waterman (1981), and Smith et al. (1981)). Practically, pairwise alignment is performed today with a limited group of software packages, most of which are freely available on the World Wide Web.

The sensitivity and specificity of the available pairwise sequence alignment algorithms continue to be assessed. Recent areas in which pairwise alignment has been further developed include methods of masking low-complexity sequences (to be discussed in Chapter 4) and theoretical models for penalizing gaps in alignments.

PITFALLS

The most common error that is committed in performing pairwise sequence alignment is when one fails to obtain the proper input sequences. In particular, it is not possible to compare a DNA to a protein sequence, or vice versa.

The optional parameters that accompany a pairwise alignment algorithm can greatly influence the results. A comparison of human RBP4 and bovine β-lactoglobulin using BLAST 2 Sequences results in no match detected if the default parameters are used.

Any two sequences can be aligned, even if they are unrelated. In some cases, two proteins that share even greater than 30% amino acid identity over a stretch of 100 amino acids are not homologous (evolutionarily related). It is always important to assess the biological significance of a sequence alignment. This may involve searching for evidence for a common cellular function, a common overall structure, or if possible a similar three-dimensional structure.

WEB RESOURCES

Pairwise sequence alignment can be performed using software packages that implement global or local alignment algorithms. In all cases, two protein or two nucleic acid sequences are directly compared.

Many websites offer freely available pairwise local alignment algorithms based upon global alignment (Table 3.4) or local alignment (Table 3.5). These sites include the NCBI's BLAST 2 Sequences, the Baylor College of Medicine (BCM) launcher, the SIM program at ExPASy, and SSEARCH at the Protein Information Resource (PIR) at Georgetown University.

A variety of commercial packages are also available, including the Genetics Computer Group (GCG) package and Vector NTI (Informax). GCG includes the GAP program (for global pairwise alignment) and BESTFIT (for local alignment). In the Unix-based version, a command such as "gap rbp.pep lactoglobulin.pep -ran=100" would run the gap global alignment algorithm on two peptide sequences and evaluate the stastistical significance of the alignment with a test of 100 randomly shuffled versions of the β-lactoglobulin sequence. A web-based interface, SeqWeb, is available.

TABLE 3-4 Global Pairwise Alignment Algorithms

Resource	Description	URL
ALIGN	At the GENESTREAM server, France	▸http://www2.igh.cnrs.fr/bin/align-guess.cgi
GAP	From the Genetics Computer Group (GCG)	▸http://www.gcg.com
Needle	From the Institut Pasteur; implements Needleman-Wunsch global alignment	▸http://bioweb.pasteur.fr/docs/EMBOSS/needle.html
Pairwise alignment (various)	From the University of Virginia (Bill Pearson)	▸http://alpha10.bioch.virginia. edu/fasta/
Pairwise	Two Sequence Alignment Tool (global and local options)	▸http://informagen.com/ Applets/Pairwise/
Pairwise Sequence Alignment	From the Baylor College of Medicine; various tools	▸http://searchlauncher.bcm. tmc.edu/
Stretcher	From the Institut Pasteur; global alignment	▸http://bioweb.pasteur.fr/docs/EMBOSS/stretcher.html
Vector NTI Suite 7	From Informax	▸http://www.informaxinc.com

Abbreviations: EMBOSS, The European Molecular Biology Open Software Suite (▸http://www.uk.embnet.org/Software/EMBOSS/); ISREC, Swiss Institute for Experimental Cancer Research (▸http://www.isrec.isb-sib.ch/).

TABLE 3-5 Local Pairwise Alignment Algorithms

Resource	Description	URL
BestFit	From the Genetics Computer Group (GCG)	▸http://www.gcg.com
BLAST 2 Sequences	At NCBI	▸http://www.ncbi.nlm.nih.gov/BLAST/
est2genome	From the Institut Pasteur; aligns expressed sequence tags to genomic DNA	▸http://bioweb.pasteur.fr/docs/EMBOSS/est2genome.html
LALIGN	Finds multiple matching subsegments in two sequences	▸http://www.ch.embnet.org/software/LALIGN_form.html
Pairwise alignment (various)	From the University of Virginia (Bill Pearson)	▸http://alpha10.bioch.virginia.edu/fasta/
Pairwise	Two Sequence Alignment Tool (global and local options)	▸http://informagen.com/Applets/Pairwise/
Pairwise Sequence Alignment	From the Baylor College of Medicine; various tools	▸http://searchlauncher.bcm.tmc.edu/
PRSS	From the University of Virginia (Bill Pearson)	▸http://fasta.bioch.virginia.edu/fasta/prss.htm
SIM	Alignment tool for protein sequences from ExPASy	▸http://www.expasy.ch/tools/sim-prot.html
SIM	SIM, gap at the Department of Computer Science, Michigan Tech	▸http://genome.cs.mtu.edu/align.html
SSEARCH	At the Protein Information Resource	▸http://pir.georgetown.edu/pirwww/
Vector NTI Suite 7	From Informax	▸http://www.informaxinc.com

Abbreviations: EMBOSS, The European Molecular Biology Open Software Suite (▸http://www.uk.embnet.org/Software/EMBOSS/); ISREC, Swiss Institute for Experimental Cancer Research (▸http://www.isrec.isb-sib.ch/).

DISCUSSION QUESTIONS

[3-1] If you want to compare any two proteins, is there any one "correct" scoring matrix to choose? Is there any way to know which scoring matrix is best to try?

[3-2] Many protein (or DNA) sequences have separate domains. (We will discuss domains in Chapter 8.) Consider a protein that has one domain that evolves rapidly and a second

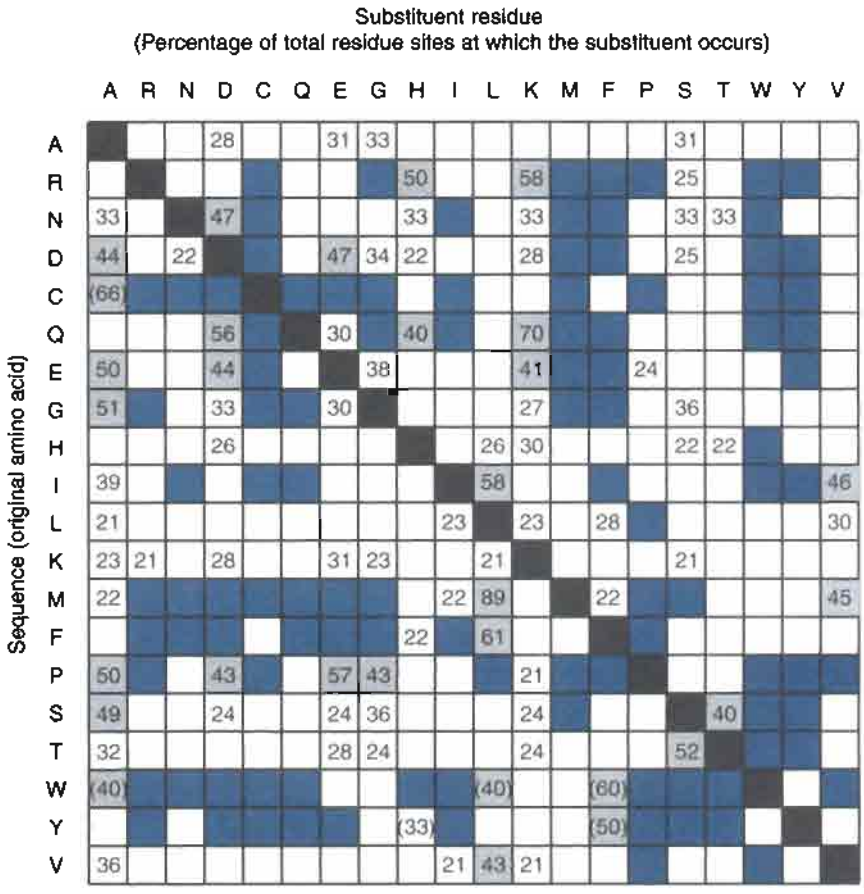


FIGURE 3.31. Substitution frequencies of globins (adapted from Zuckerkandl and Pauling, 1965, p. 118). Amino acids are presented alphabetically according to the three letter abbreviations. The rows correspond to an original amino acid in an alignment of several dozen hemoglobin and myoglobin protein sequences from human, other primates, horse, cattle, pig, lamprey, and carp. Numbers represent the percentages of residue sites at which a given substitution occurs. For example, a glycine substitution was observed to occur in 33% of all the alanine sites. Substitutions that were never observed to occur are indicated by squares colored red. Rarely occurring substitutions (percentages <20%) are indicated by empty white squares (numerical values are not given). "Very conservative" substitutions (percentages ≥40%) are in boxes shaded gray. For example, in 89% of the sites containing a methionine, lecuine was also observed to be present. Identities are indicated by black solid squares. Values in parentheses indicate a very small available sample size, suggesting that conclusions about those data should be made cautiously. Used with permission.

domain that evolves slowly. In doing a pairwise alignment with another protein (or DNA) sequence, would you use two separate alignments with scoring matrices such as PAM40 and PAM250 or would you select one "intermediate" matrix? Why?

[3-3] Thirteen years before Margaret Dayhoff and colleagues published a protein atlas with scoring matrices, Emile Zuckerkandl and Linus Pauling (1965) produced a scoring matrix for several dozen available globin sequences (Figure 3.31). The rows (y-axis) of this figure show the original globin amino acid, and the columns show substitutions that were observed to occur. Numerical values are entered in cells for which the substitutions occur in at least 20% of the sites. Note that for cells shaded red, these amino acid substitutions were never observed, while for cells shaded gray the amino acid substitutions were defined as very conservative.

How do the data in this matrix compare to those described by Dayhoff and colleagues? Which substitutions occur most rarely, and which most frequently? How would you go about filling in this table today?

Joshua Lederberg helped Zuckerkandl and Pauling (1965) make this matrix. They used an IBM 7090 computer, one of the first commercial computers based on transistor technology. The computer cost about $3 million. Its memory consisted of 32,768 binary words or about 131,000 bytes. (To read about Lederberg's Nobel Prize from 1958, see ▸ http://www.nobel.se/medicine/laureates/1958/index.html.)

PROBLEMS

[3-1] Viral reverse transcriptases, such as the *pol* gene product encoded by HIV-1, have human homologs. The GenBank accession number for HIV-1 reverse transcriptase is NP.057849. (Use Entrez to confirm this is the correct accession number.) A search of Entrez reveals many human viral-related gene products including a retrovirus-related Pol polyprotein of 874 amino acid residues (P10266). Use BLAST 2 Sequences to perform a pairwise alignment using the blastp program.

The default conditions for this search include the use of the BLOSUM62 scoring matrix. The results are shown in Figure 3.32 and Table 3.6. The expect value is about

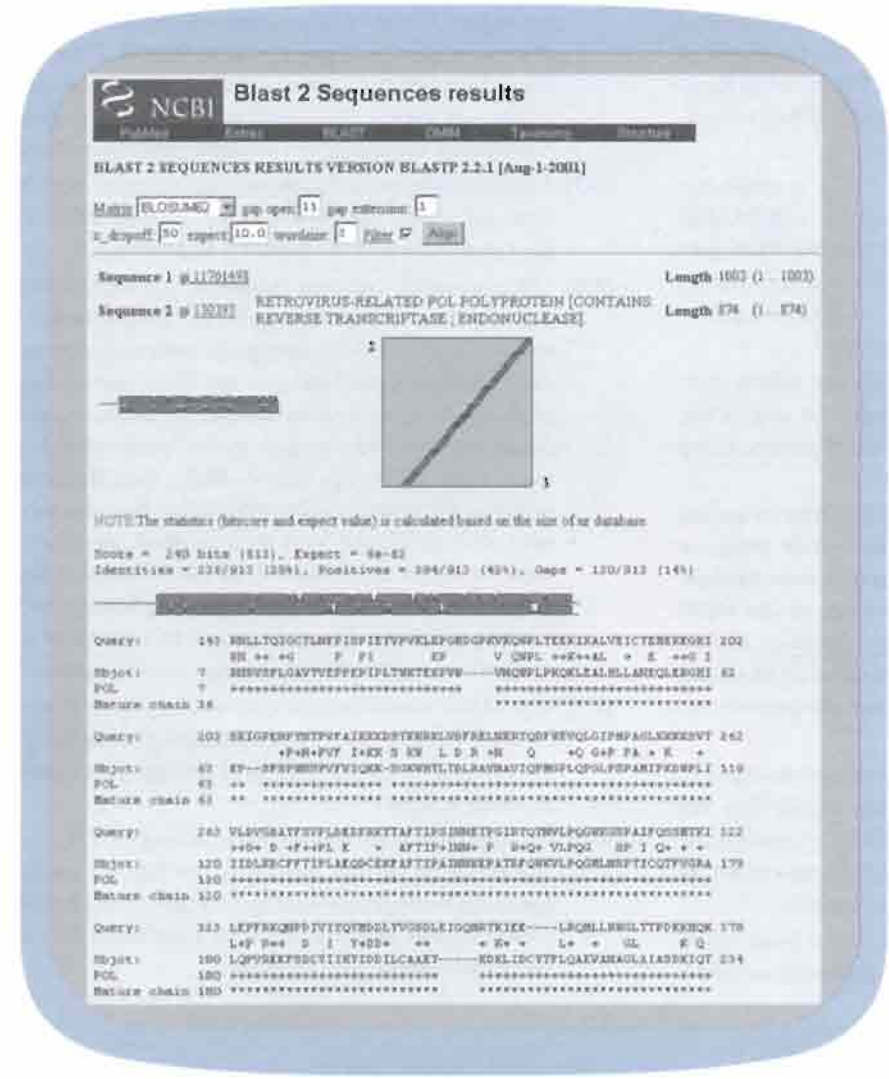


FIGURE 3.32. *Partial output of a pairwise alignment between HIV-1 pol (NP.057849) and a human reverse transcriptase (P10266). Overall these proteins share 25% amino acid identity.*

TABLE 3-6 Effects of Changing Scoring Matrices on Pairwise Alignment of HIV-1 *pol* and Related Human Protein

Scoring Matrix	Score	Expect Value	Identities	Positives	Gap Penalties	Gaps
BLOSUM50						
BLOSUM62	240 bits (613)	1×10^{-61}	236/913 (25%)	394/913 (43%)	GO = 11, GE = 1	130/913 (14%)
BLOSUM90						
PAM30						
PAM70						
PAM250						

Note: GO denotes gap opening; GE is gap extension.

1×10^{-61}, indicating that the proteins are closely related even though they share only 25% identity over a span of 919 amino acids. Fill in the table using the BLOSUM62, BLOSUM50, and BLOSUM90 scoring matrices. What is the effect of changing the search parameters?

[3-2] Next perform pairwise alignments of the proteins described in problem 3-1 using the PAM30, PAM70, and PAM250 matrices. What are the expect values? What span of amino acid residues is aligned? Are the search results using different PAM matrices similar or different to the results of using different BLOSUM matrices?

[3-3] Perform a local pairwise alignment between RBP and β-lactoglobulin using BLAST 2 Sequences (or any of the programs listed in Table 3.5). Repeat the alignment using lower gap penalties. What is the result?

[3-4] Compare modern human mitochondrial DNA to extinct Neanderthal DNA. First obtain the nucleotide sequence of a mitochondrially encoded gene, cytochrome oxidase. (Begin by searching the taxonomy division of the NCBI website, ▸ http://www.ncbi.nlm.nih.gov/taxonomy, and select "extinct organisms" to find Neanderthal DNA.) Next, perform pairwise alignments and record the percent nucleotide identities.

[3-5] We have seen that some gene families change slowly (e.g., GAPDH in Fig. 3.8) while others change rapidly (e.g., see Box 3.3). How can you determine whether the cytochrome oxidase gene you studied in problem 3.4 changes relatively rapidly or slowly? Try using BLAST 2 Sequences.

[3-6] Aphrodisin and odorant-binding protein are both examples of lipocalins. First obtain the accession numbers for rodent forms of these proteins, and then perform a pairwise sequence alignment. (Use BLAST 2 Sequences.) Record the percent amino acid identity, the percent similarities, the expect values, and bit scores. Which metric is most useful in helping you evaluate their relatedness?

[3-7] Are crocodiles more closely related to turtles or to birds? First, try to find the answer using PubMed. Then select a gene or protein and use BLAST 2 Sequences.

[3-8] The coelacanth *Latimeria chalumnae* is a lobe-finned fish that has been called a "living fossil." Long thought to be extinct for at least 90 million years, several specimens have now been discovered lurking in the ocean. Incredibly, some phylogenetic analyses of mitochondrial DNA sequences indicate that the coelacanth is more closely related to humans than to herrings (Lewin, 2001). Find the accession numbers for some mitochrondrial DNA from human, herring, and coelacanth, and then perform pairwise alignments to decide if you agree. *Hint:* Use PubMed to find the genus and species name of an organism; herring is *Clupea harengus*. Next use this species name in a search of Entrez nucleotides, such as "*Clupea harengus* mitochondrion."

[3-9] The PAM1 matrix (Fig. 3.11) is nonreciprocal: The probability of changing an amino acid such as alanine to arginine is not equal to the probability of changing an arginine to an alanine. Why?

[3-10] Is a hippopotamus more closely related to a pig or to a whale? To answer this question, first find the protein sequence of hemoglobin from each of those three organisms. Next, perform pairwise sequence alignments and record the percent amino acid identities.

SELF-TEST QUIZ

[3-1] Match the following amino acids with their single-letter codes:

Asparagine	Q
Glutamine	W
Tryptophan	Y
Tyrosine	N
Phenylalanine	F

[3-2] Orthologs are defined as:

(a) Homologous sequences in different species that share an ancestral gene
(b) Homologous sequences that share little amino acid identity but share great structural similarity
(c) Homologous sequences in the same species that arose through gene duplication
(d) Homologous sequences in the same species which have similar and often redundant functions

[3-3] Which of the following amino acids is least mutable according to the PAM scoring matrix?

(a) Alanine
(b) Glutamine
(c) Methionine
(d) Cysteine

[3-4] The PAM250 matrix is defined as having an evolutionary divergence in which what percentage of amino acids between two homologous sequences have changed over time?

(a) 1%
(b) 20%
(c) 80%
(d) 250%

[3-5] Which of the following sentences best describes the difference between a global alignment and a local alignment between two sequences?

(a) Global alignment is usually used for DNA sequences, while local alignment is usually used for protein sequences.

(b) Global alignment has gaps, while local alignment does not have gaps.
(c) Global alignment finds the global maximum, while local alignment finds the local maximum.
(d) Global alignment aligns the whole sequence, while local alignment finds the best subsequence that aligns.

[3-6] You have two distantly related proteins. Which BLOSUM or PAM matrix is best to use to compare them?
(a) BLOSUM45 or PAM250
(b) BLOSUM45 or PAM1
(c) BLOSUM80 or PAM250
(d) BLOSUM80 or PAM1

[3-7] How does the BLOSUM scoring matrix differ most notably from the PAM scoring matrix?
(a) It is best used for aligning very closely related proteins.
(b) It is based on global multiple alignments from closely related proteins.
(c) It is based on local multiple alignments from distantly related proteins.
(d) It combines local and global alignment information.

[3-8] True or false: Two proteins that share 30% amino acid identity are 30% homologous.

[3-9] A global alignment algorithm (such as the Needleman-Wunsch algorithm) is guaranteed to find an optimal alignment. Such an algorithm:
(a) Puts the two proteins being compared into a matrix and finds the optimal score by exhaustively searching every possible combination of alignments
(b) Puts the two proteins being compared into a matrix and finds the optimal score by iterative recursions
(c) Puts the two proteins being compared into a matrix and finds the optimal alignment by finding optimal subpaths that define the best alignment(s)
(d) Can be used for proteins but not for DNA sequences

[3-10] In a database search or in a pairwise alignment, sensitivity is defined as:
(a) The ability of a search algorithm to find true positives (i.e., homologous sequences) and to avoid false positives (i.e., unrelated sequences having high similarity scores)
(b) The ability of a search algorithm to find true positives (i.e., homologous sequences) and to avoid false positives (i.e., homologous sequences that are not reported)
(c) The ability of a search algorithm to find true positives (i.e., homologous sequences) and to avoid false negatives (i.e., unrelated sequences having high similarity scores)
(d) The ability of a search algorithm to find true positives (i.e., homologous sequences) and to avoid false negatives (i.e., homologous sequences that are not reported)

SUGGESTED READING

We introduced this chapter with the concept of homology, an often misused term. A one-page article by Reeck et al. (1987) provides authoritative, standard definitions of the terms homology and similarity. A more recent discussion of homology in relation to phylogeny is provided by Tautz (1998).

All of the papers describing sequence alignment consider the divergence of two homologous sequences in the context of a model of molecular evolution. Russell F. Doolittle (1981) has written a clear, thoughtful overview of sequence alignment. William Pearson (1996) has reviewed sequence alignment. He provides descriptions of the statistics of similarity scores, sensitivity and selectivity, and search programs such as Smith-Waterman and FASTA. Other reviews of pairwise alignment include short articles by Altschul (1998) and Brenner (1998) in *Trends Guide to Bioinformatics*.

For studies of pairwise sequence alignment algorithms, an important historical starting point is the 1978 book by Margaret O. Dayhoff and colleagues (Dayhoff, 1978) Most of this book consists of an atlas of protein sequences with accompanying phylogenetic reconstructions. Chapter 22 introduces the concept of accepted point mutations, while Chapter 23 describes various PAM matrices. By the early 1990s, when far more protein sequence data were available, Steven and Jorja Henikoff (1992) described the BLOSUM matrices. This article provides an excellent technical introduction to the use of scoring matrices, usefully contrasting the performance of PAM and BLOSUM matrices. Later (in Chapters 4 and 5) we will use these matrices extensively in database searching.

The algorithms originally describing global alignment are presented technically by Needleman and Wunsch (1970) and later local alignment algorithms were introduced by Smith and Waterman (1981) and Smith, Waterman, and Fitch (1981). The problem of both sensitivity (the ability to identify distantly related sequences) and selectivity (the avoidance of unrelated sequences) of pairwise alignments was addressed by Pearson and Lipman in a 1988 paper introducing the FASTA program.

More recent articles address technical aspects of sequence-scoring statistics. Marco Pagni and C. Victor Jongeneel (2001) of the Swiss Institute of Bioinformatics provide an excellent overview. This includes a discussion of BLAST scoring statistics that is relevant to Chapters 4 and 5.

Finally, Steven Brenner, Cyrus Chothia, and Tim Hubbard (1998) have compared several pairwise sequence methods. This article is highly recommended as a way to learn how different algorithms can be assessed (we will see similar approaches for multiple sequence alignment in Chapter 10, for example). Reading this paper can help to show why statistical scores are more effective than other search parameters such as raw scores or percent identity in interpreting pairwise alignment results.

REFERENCES

Altschul, S. F. Fundamentals of database searching. *Bioinformatics: A Trends Guide* **1998,** 7–9 (1998).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

Anfinsen, C. *The Molecular Basis of Evolution.* John Wiley & Sons, Inc., New York, 1959.

Brenner, S. E. Practical database searching. *Bioinformatics: A Trends Guide* **1998,** 9–12 (1998).

Brenner, S. E., Chothia, C., and Hubbard, T. J. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA* **95,** 6073–6078 (1998).

Chothia, C., and Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *Embo J.* **5,** 823–826 (1986).

Dayhoff, M. O. *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Silver Spring, MD, 1978.

Doolittle, R. F. Similar amino acid sequences: Chance or common ancestry? *Science* **214,** 149–159 (1981).

Doolittle, R. F. *OF URFS AND ORFS: A Primer on How to Analyze Derived Amino Acid Sequences.* University of Science Books, Mill Valley, CA, 1987.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. *Biological Sequence Analysis.* Cambridge University Press, Cambridge, 2000.

Gonnet, G. H., Cohen, M. A., and Benner, S. A. Exhaustive matching of the entire protein sequence database. *Science* **256,** 1443–1445 (1992).

Gotoh, O. An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162,** 705–708 (1982).

Henikoff, S., and Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919 (1992).

Henikoff, J. G., and Henikoff, S. Blocks database and its applications. *Methods Enzymol.* **266,** 88–105 (1996).

Lewin, R. A. Why rename things? *Nature* **410,** 637 (2001).

Myers, E. W., and Miller, W. Optimal alignments in linear space. *Comput. Appl. Biosci.* **4,** 11–17 (1988).

Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453 (1970).

Newcomer, M. E., et al. The three-dimensional structure of retinol-binding protein. *EMBO J.* **3,** 1451–1454 (1984).

Pagni, M., and Jongeneel, C. V. Making sense of score statistics for sequence alignments. *Brief Bioinform.* **2,** 51–67 (2001).

Pearson, W. R. Effective protein sequence comparison. *Methods Enzymol.* **266,** 227–258 (1996).

Pearson, W. R., and Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448 (1988).

Reeck, G. R., et al. "Homology" in proteins and nucleic acids: A terminology muddle and a way out of it. *Cell* **50,** 667 (1987).

Sellers, P. H. On the theory and computation of evolutionary distances. *SIAM J. Appl. Math.* **26,** 787–793 (1974).

Smith, T. F., and Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).

Smith, T. F., Waterman, M. S., and Fitch, W. M. Comparative biosequence metrics. *J. Mol. Evol.* **18,** 38–46 (1981).

Tatusova, T. A., and Madden, T. L. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett.* **174,** 247–250 (1999).

Tautz, D. Evolutionary biology. Debatable homologies. *Nature* **395,** 17, 19 (1998).

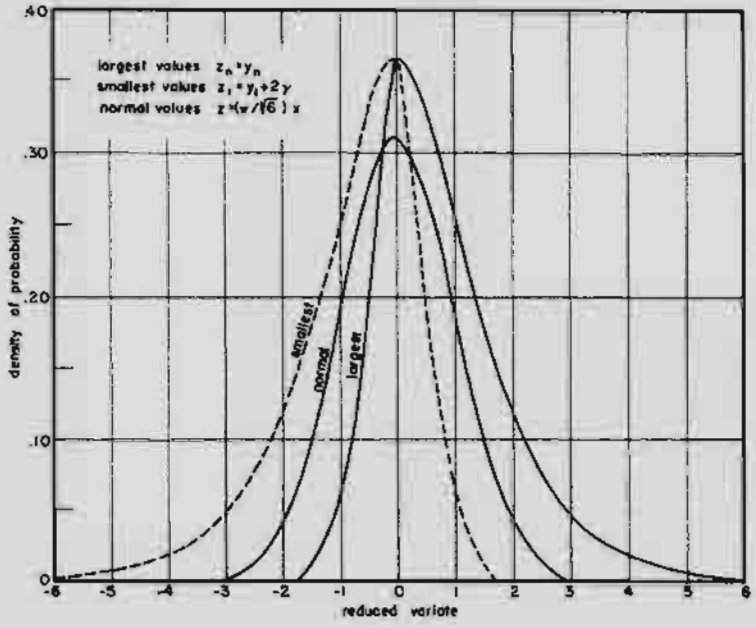
Zuckerkandl, E., and Pauling, L. Evolutionary divergence and convergence in proteins. In: V. Bryson and H. J. Vogel (eds.), *Evolving Genes and Proteins,* Academic Press, New York, 1965, pp. 97–166.

This Page Intentionally Left Blank

of the variate, the first double exponential distribution has larger (smaller) densities than the normal one. The opposite is true for the second double exponential distribution.

Table 5.2.7. Selected Probabilities for Normal and Largest Values

Value	Reduced Variate		Probabilities		Return Periods	
	Largest	*Normal*	*Largest*	*Normal*	*Largest*	*Normal*
$\tilde{x} - \sigma$	−.70533	−1	.13206	.15866	7.57	6.30
$\tilde{x} + \sigma$	1.85977	1	.85581	.84134	6.93	6.30
$\tilde{x} \pm \sigma$	—	—	.72375	.68268	—	—
$\tilde{x} - 2\sigma$	−1.98788	−2	.00068	.02275	1480.	43.96
$\tilde{x} + 2\sigma$	3.14232	2	.95773	.97725	23.7	43.96
$\tilde{x} \pm 2\sigma$	—	—	.95705	.95450	—	—
$\tilde{x} - 3\sigma$	−3.27043	−3	$3.7 \cdot 10^{-12}$	.00135	$.27 \cdot 10^{11}$	741
$\tilde{x} + 3\sigma$	4.42486	3	.98810	.99865	84.01	741
$\tilde{x} \pm 3\sigma$	—	—	.98810	.99730	—	—



Graph 5.2.7(1). Extreme and Normal Distributions

In Graph 5.2.7(2) the probabilities of the largest and the smallest values and the normal probabilities for the same mean and standard deviation

Chapter 4 describes the principal database search tool, BLAST. While BLAST was first described by Altschul et al. in 1990, the statistical interpretation of the scores you get in a BLAST search are based on mathematical models developed by the 1950s. In many instances, the distribution of values in a population assumes a normal (Gaussian) distribution, as shown in this figure (see curve labeled "normal"). However, for a wide variety of natural phenomena the distribution of extreme values is not *normal. Such is the case for database searches in which you search with a protein or DNA sequence of interest (the query) against a large database, as will be described in this chapter. The maximum scores fit an extreme value distribution (EVD) rather than a normal distribution. In 1958 Emil Gumbel described the statistical basis of the EVD in his book* Statistics of Extremes. *This figure (Gumbel, 1958, p. 180) shows the EVD. Note that for the curve marked "largest" the tail is skewed to the right. Also, as shown in the table, for a normal distribution values that are up to three standard deviations above the mean occupy 99.865% of the area under the curve, while for the EVD values up to three standard deviations occupy only 98.810%. In other words, the EVD is characterized by a larger area under the curve at the extreme right portion of the plot. We will see how this analysis is applied to BLAST search results to let you assess whether a query sequence is significantly related to a match in the database. Used with permission.*

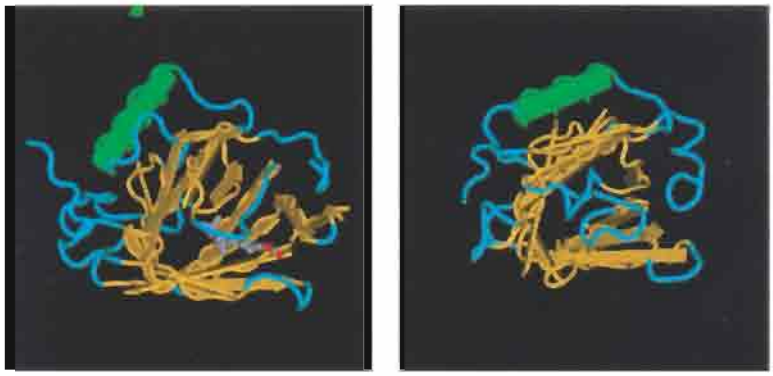


FIGURE 3.1 (page 42). *Three-dimensional structures of two lipocalins: bovine retinol-binding protein (RBP; left panel), bovine β-lactoglobulin (right panel). The images were generated with the program Cn3D (see Chapter 9). These proteins are homologous (descended from a common ancestor), and they share very similar 3-dimensional structures consisting of a binding pocket for a ligand and eight antiparallel beta sheets. However, pairwise alignment of these proteins' amino acid sequences reveals that the proteins share very limited amino acid identity. The accession numbers are MMDB Id: 934 PDB Id: 1FEM (RBP), MMDB Id: 11969 PDB Id: 1BSQ (β-lactoglobulin).*

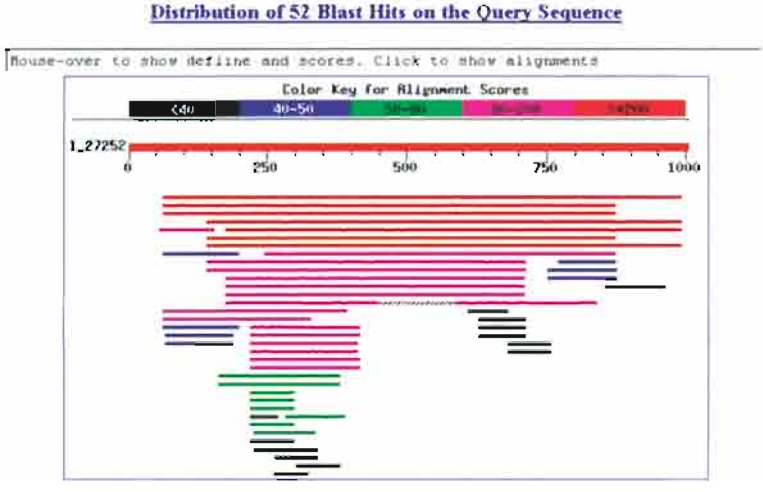


FIGURE 4.28 (page 119). *Graphical output of a blastp search using HIV-1 pol protein to search for matches against human proteins.*

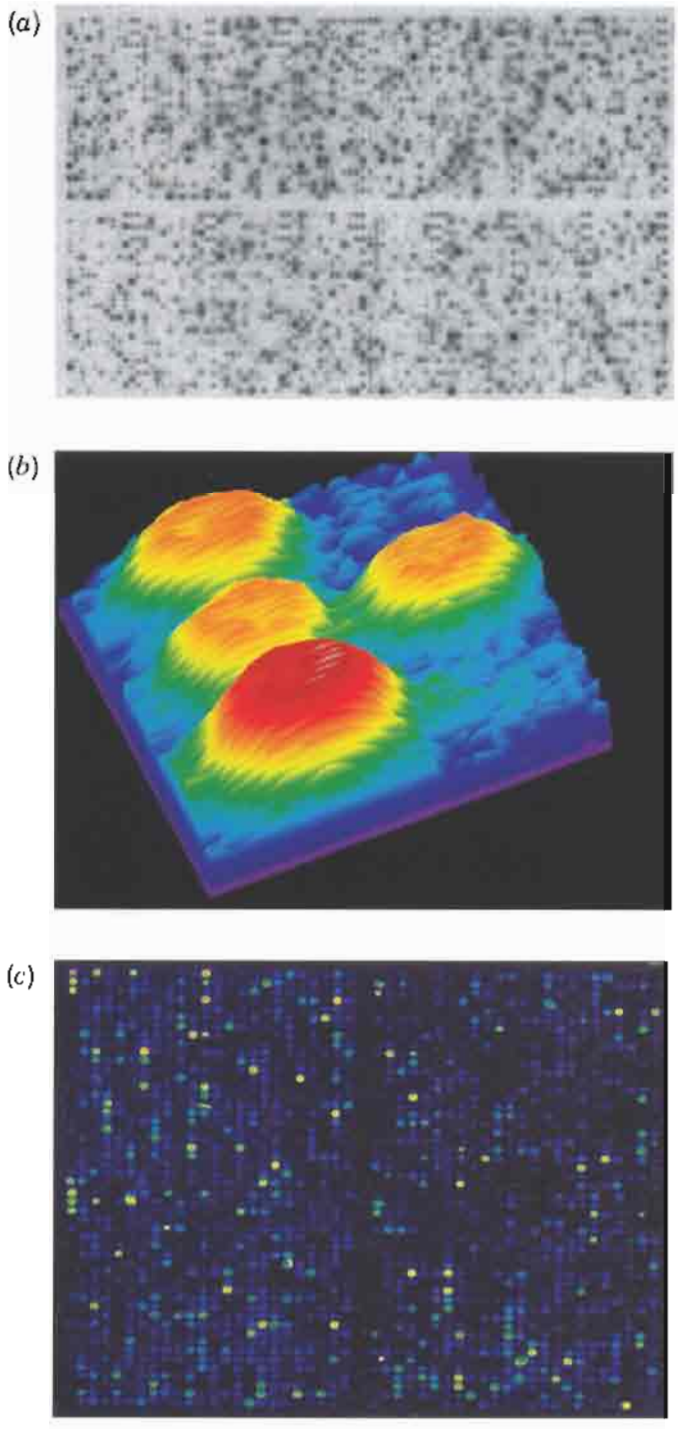
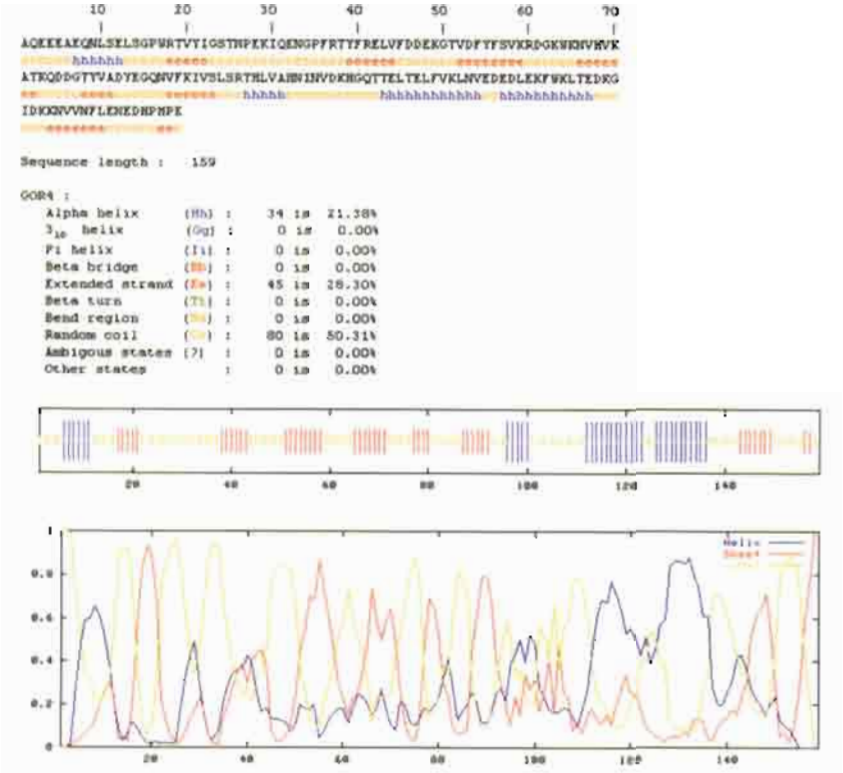


FIGURE 6.20 (page 181). *GeneFilter from Research Genetics is probed with [^{32}P]cDNA derived from the hippocampus of a postmortem brain of an individual with Down Syndrome. (a) There are 5000 cDNAs spotted on the array. The pattern in which genes are represented on any array must be randomized. (b) Six of the signals are visualized using NIH Image software. Image analysis software must define the properties of each signal, including the likelihood that an intense signal (lower left) will "bleed" onto a weak signal (lower right). (c) A microarray from NEN Perkin Elmer (MICROMAX, representing 2400 genes) was probed with the same Rett Syndrome and control brain samples used in Figure 6.19. This technology employs cDNA samples that are fluorescently labeled in a competitive hybridization.*

(a)



(b)

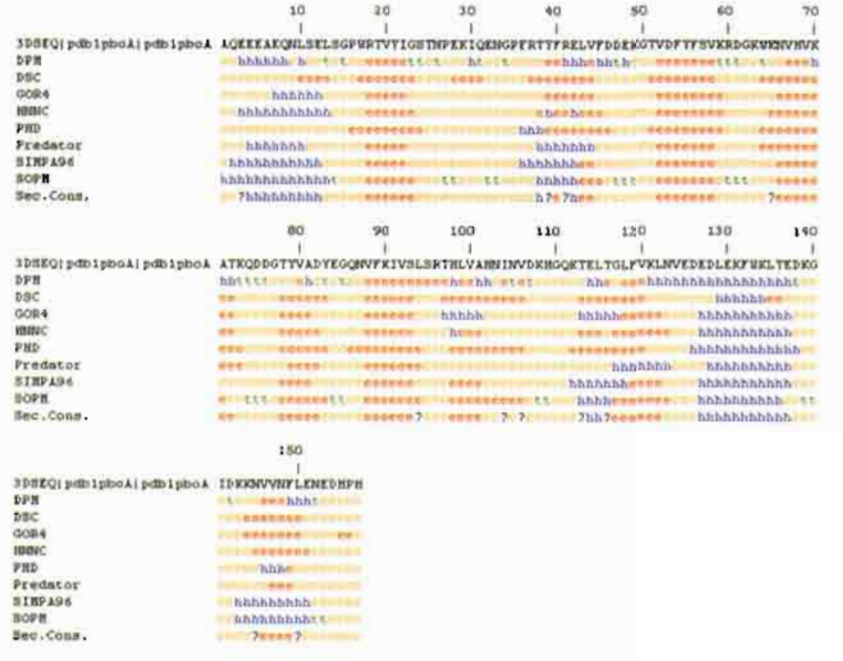


FIGURE 9.3 (page 277). *A variety of web servers offer secondary structure prediction. The sequence of a lipocalin (bovine odorant-binding protein, accession P07435) was entered in the Network Protein Sequence Analysis site of the Pôle Bio-Informatique Lyonnais. (a) Secondary structure predictions such as alpha helices are shown as well as (b) the combined results of nine prediction algorithms. Note that these algorithms offer slightly differing predictions. The letters c, e, h, and t (panel a) are defined in panel b.*

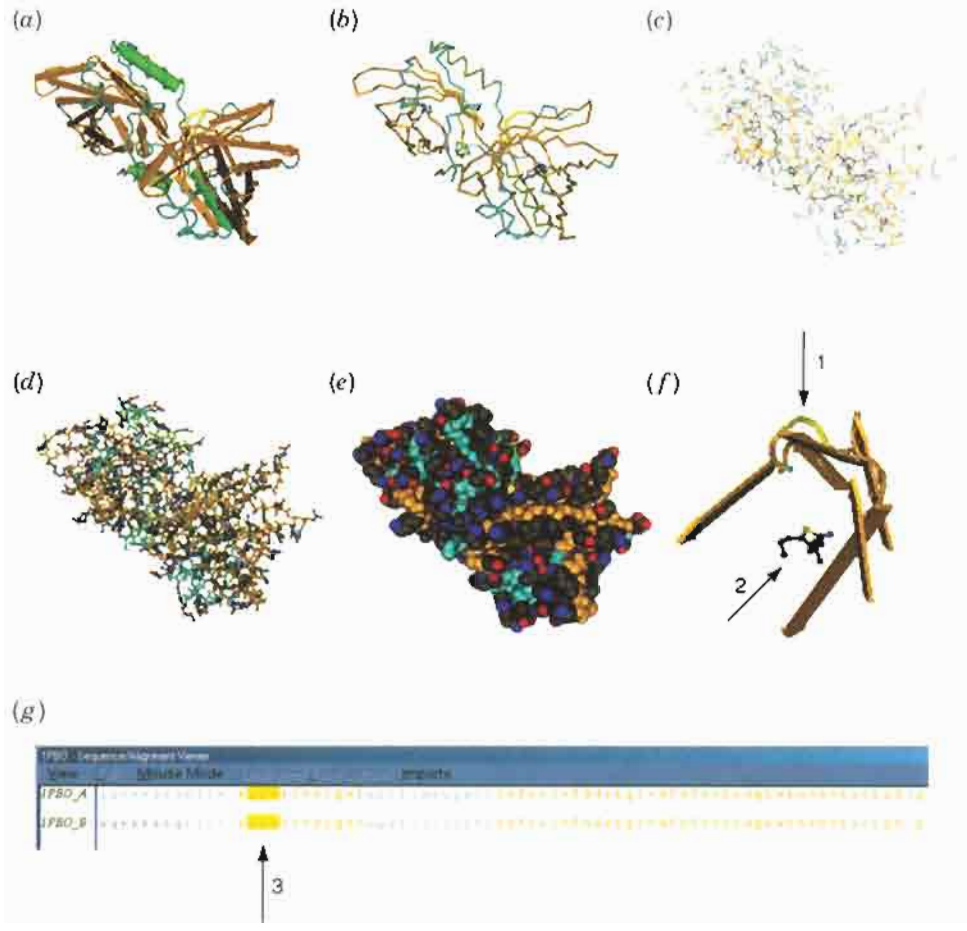


FIGURE 9.16 (page 291). *The NCBI Cn3D viewer shows the secondary structure of the odorant-binding protein (1PBO), an eight-stranded beta barrel, in a variety of formats: (a) worms, (b) tubes, (c) wire, (d) ball and stick, and (e) spacefill. (f) It is possible to rotate the image and zoom on any region of the protein. (g) The primary amino acid sequence is shown in a Sequence/Alignment Viewer. By clicking on a sequence consisting of the conserved GWX motif (arrow 3), it can be highlighted in the protein (arrow 1). There are a variety of other visualization options, such as an image of the ligand (an odorant) in the binding pocket of the protein (arrow 2).*

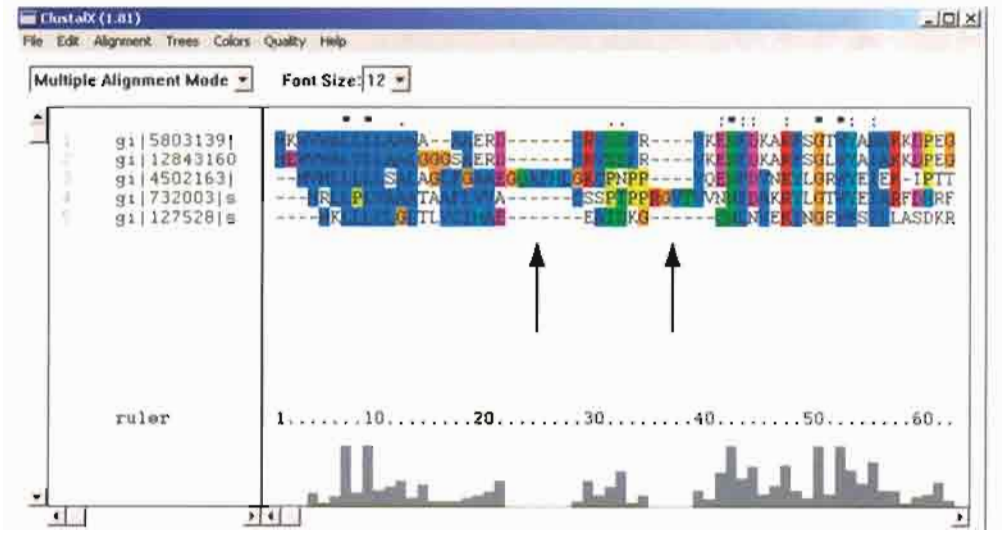


FIGURE 10.24 (page 344). *The CLUSTAL X program creates multiple sequence alignments. Five lipocalin protein sequences were obtained from the NCBI Entrez site in the FASTA format, copied into a word document, and saved as a text file (ASCII format) (see Figure 10.1 for proteins). This file was then input into CLUSTAL X with the "load sequences" command. The sequences are initially entered into CLUSTAL X without gap insertions. Using the "Do Complete Alignment" command under the Alignment pull-down menu, a multiple sequence alignment is produced. This can be edited manually. Asterisks above the aligned columns indicate positions of 100% amino acid identity. The histogram at the bottom of the display box indicates the relative amount of sequence conservation across the alignment. CLUSTAL X alignments can be saved for phylogenetic analysis using other software such as PAUP (Chapter 11).*

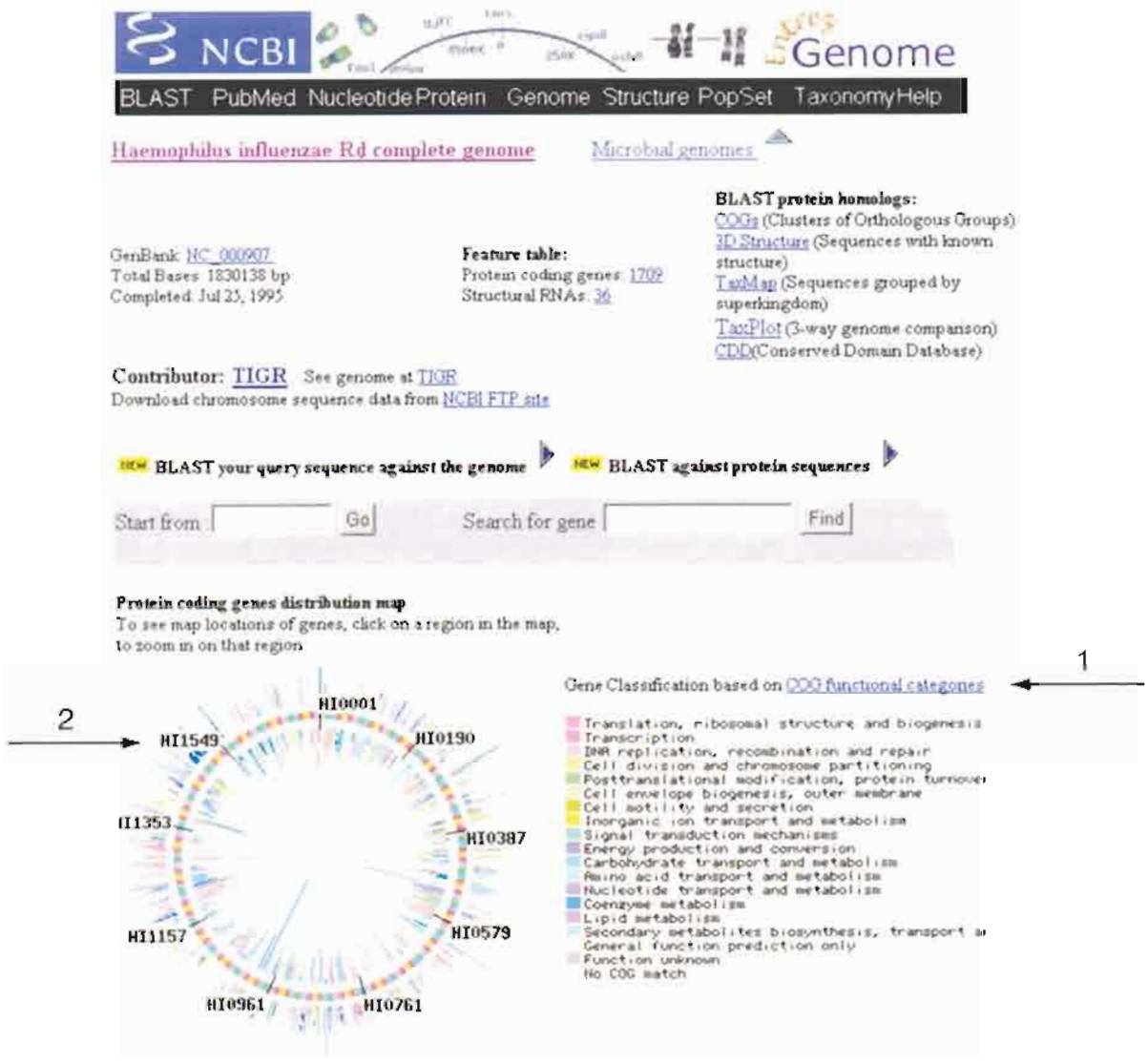


FIGURE 12.9 (page 411). *Entrez Genome record for* Haemophilus influenzae *Rd, the first free-living organism for which the complete genomic sequence was determined. This record is obtained from the Entrez Genomes resource by clicking "bacteria" on the left sidebar. The top of this entry includes information such as the accession number and the size of the genome. The entire nucleotide sequence is downloadable here. At top right are several resources for studying the 1709 proteins encoded by this genome. An* H. influenzae*-specific BLAST search is available here. The main part of the entry consists of a color-coded circular representation of the genome, along with a functional classification (arrow 1) based on Clusters of Orthologous Genes (COGs; see below). The circular map is clickable (see Fig. 12.10), showing detailed information on the genes and proteins in this genome. At the bottom, the genetic code used by this bacterium is provided, as well as the taxonomic lineage. The record also contains literature references (not shown) including the initial report of this genomic sequence by Fleischmann et al. (1995) at The Institute for Genomic Research.*

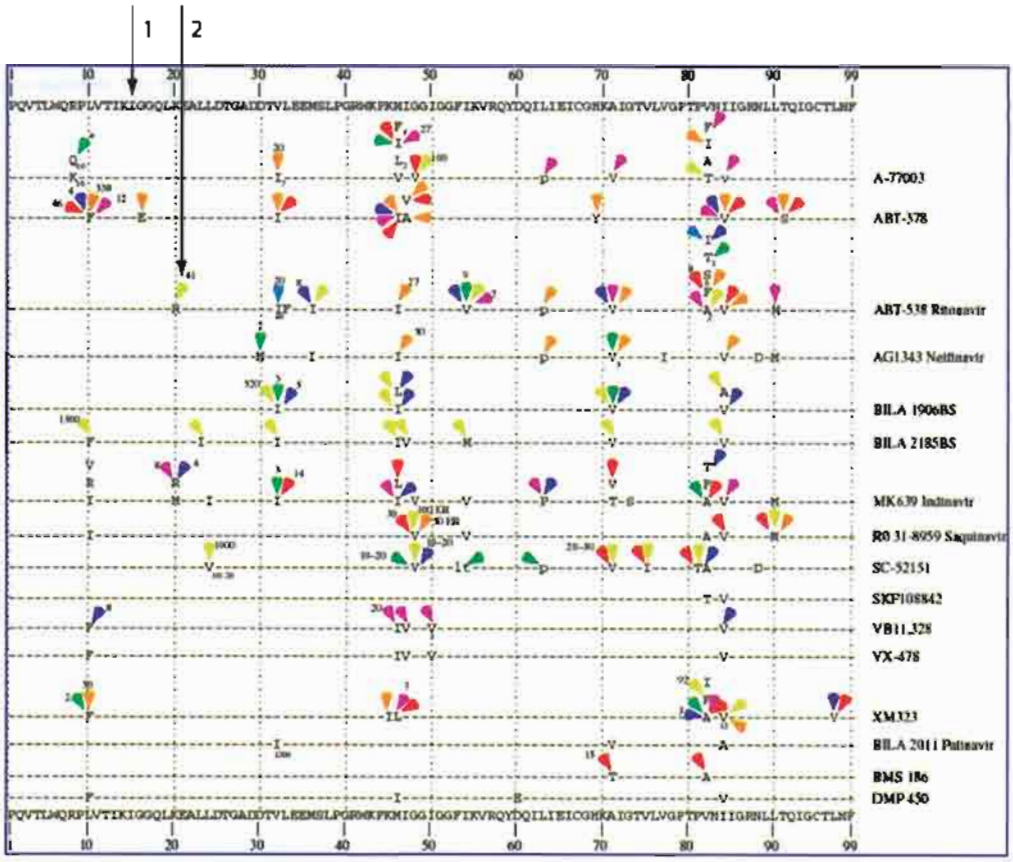


FIGURE 13.16 (page 456). *The LANL web site offers a map of HIV-1 protease mutations versus drugs. Each row represents a drug (labeled at right). The wild-type (strain HXB2) HIV-1 protease sequence is listed at top and bottom (arrow 1). Dashes indicate wild-type amino acid positions, while mutations that confer resistance to the drug are indicated. An example of a K-to-R (lysine to arginine) mutation is indicated (arrow 2). The small number (41) indicates the "fold resistance" of that particular mutation. Mutations that have a colored shape pointing to them are also part of a synergistic combination of mutations. By clicking on a mutation (arrow 2), the map links to a detailed report of the effects of that mutation.*

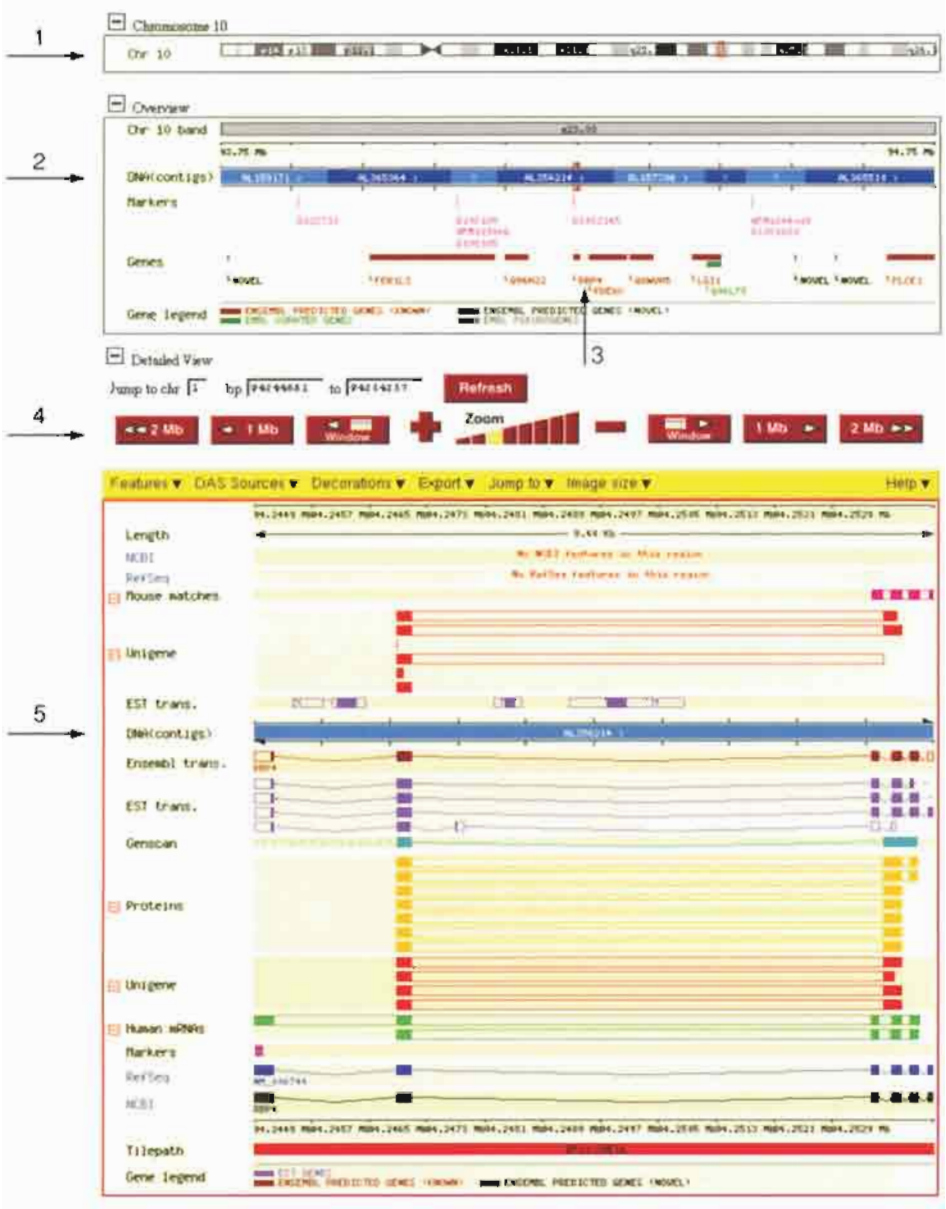
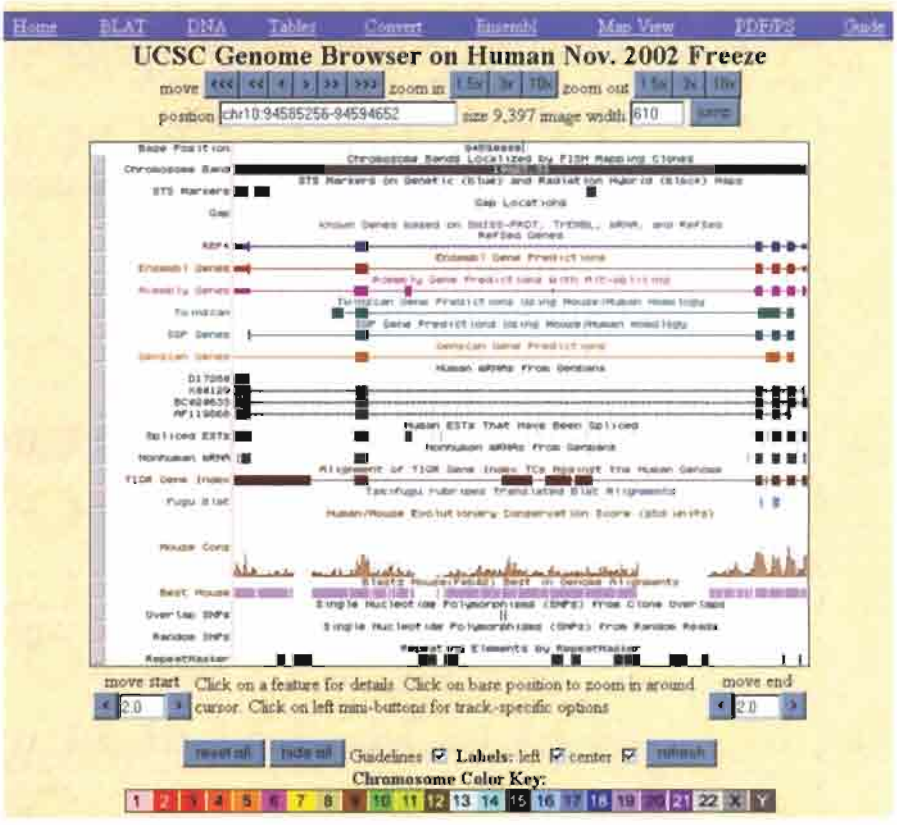


FIGURE 17.7 (page 617). *The Ensembl ContigView (linked from the* RBP4 *gene report, Figure 17.5) shows an ideogram of chromosome 10 (arrow 1), including the position of the* RBP4 *gene (boxed). The overview section includes an overview of contigs in the 10q23.33 region (arrow 2), and indicates annotated genes including* RBP4 *(arrow 3). The detailed view section can be zoomed and scrolled (arrow 4). It includes a schematic view of a contig (arrow 5) with the annotations on the top strand above it, and the annotations on the bottom strand below (including* RBP4 *links to ESTs, markers, and predicted transcripts).*



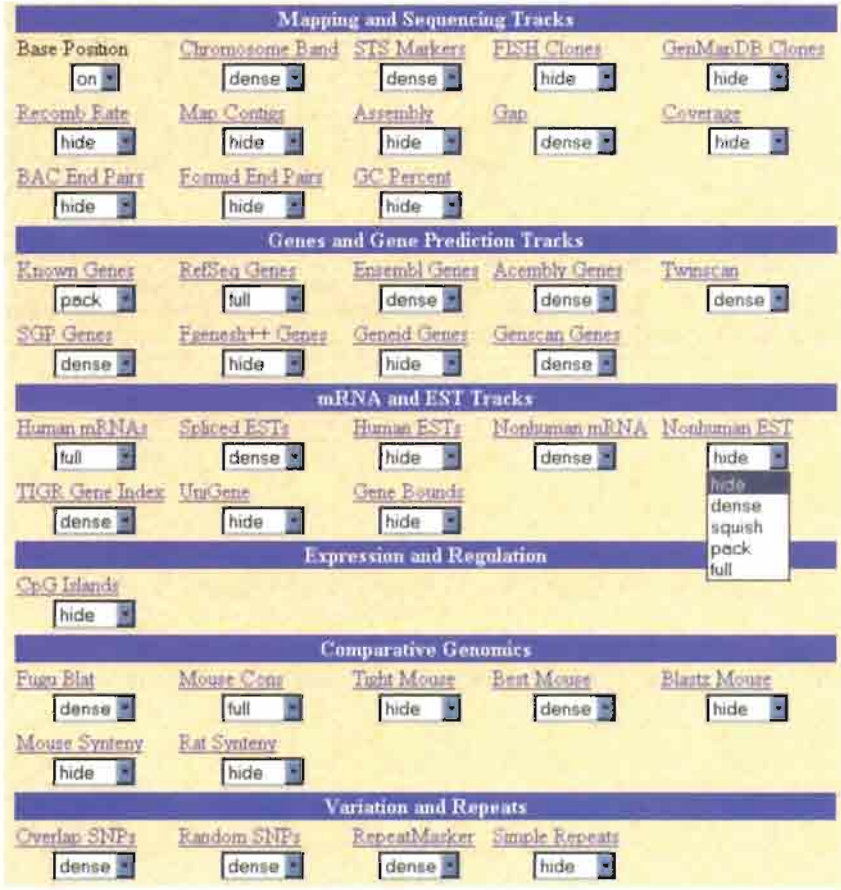


FIGURE 17.12 (page 622). *The human genome browser at UCSC (http://genome.ucsc.edu) (Kent et al., 2002). At top, the user can scroll across a chromosome. In the middle, genome annotations are displayed. At the bottom, a variety of track controls allow the display to be dynamically modified with dozens of types of annotation such as repetitive DNA, comparative genomics, and gene expression data.*

4

Basic Local Alignment Search Tool (BLAST)

INTRODUCTION

Basic Local Alignment Search Tool (BLAST) is the main NCBI tool for comparing a protein or DNA sequence to other sequences in various databases (Altschul, 1990, 1997). BLAST searching is one of the fundamental ways of learning about a protein or gene: The search reveals what related sequences are present in the same organism and other organisms. The NCBI website includes several excellent resources for learning about BLAST.

NCBI resources include a tutorial and a course that can be accessed through the sidebar on the main BLAST page (▸http://www.ncbi.nlm.nih.gov/BLAST/).

In Chapter 3, we described how to perform a pairwise sequence alignment between two protein or nucleotide sequences. BLAST searching allows the user to select one sequence (termed the *query*) and perform pairwise sequence alignments between the query and an entire database (termed the *target*). Typically, this means that millions of alignments are analyzed in a BLAST search, and only the most closely related matches are returned. The Needleman-Wunsch (1970) and Smith-Waterman (1981) global and local alignment algorithms find optimal pairwise alignments, but we cannot use them for database searches because they are too computationally intensive (it would take too much time to search a query against a large database using those algorithms). BLAST offers both speed and sensitivity, as

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

described in this chapter. It also offers convenient accessibility on the World Wide Web.

The BLAST programs allow a DNA or protein query sequence to be compared to a DNA or protein database. A DNA sequence can be converted into six potential proteins (see below), and the BLAST algorithms include strategies to compare protein sequences to dynamically translated DNA databases or vice versa. BLAST searching thus has a wide variety of uses. These include:

- *Determining what orthologs and paralogs are known for a particular protein or nucleic acid sequence.* Besides RBP, what other lipocalins are known? When a new bacterial genome is sequenced and several thousand proteins are identified, how many of these proteins are paralogous? How many of the predicted genes have no significantly related matches in GenBank?
- *Determining what proteins or genes are present in a particular organism.* Are there any lipocalins such as RBP in plants? Are there any reverse transcriptase genes (such as HIV-1 *pol* gene) in fish?
- *Determining the identity of a DNA or protein sequence.* For example, you may perform a subtractive hybridization experiment or a microarray experiment and learn that a particular DNA sequence is dramatically regulated under the experimental conditions that you are using. This DNA sequence may be searched against a protein database to learn what proteins are most related to the protein encoded by your DNA sequence.
- *Discovering new genes.* For example, a BLAST search of genomic DNA may reveal that the DNA encodes a protein that has not been described before. In Chapter 5, we will show how BLAST searching can be used to find novel, previously uncharacterized genes.
- *Determining what variants have been described for a particular gene or protein.* For example, many viruses are extremely mutable; what HIV-1 *pol* variants are known?
- *Investigating expressed sequence tags that may exhibit alternative splicing.* There is an EST database that can be explored by BLAST searching. Indeed, there are dozens of specialized databases that can be searched. For example, specialized databases consist of sequences from a specific organism, a tissue type, a chromosome, a type of DNA (such as untranslated regions), or a functional class of proteins.
- *Exploring amino acid residues that are important in the function and/or structure of a protein.* The results of a BLAST search can be multiply aligned (Chapter 10) to reveal conserved residues such as cysteines that are likely to have important biological roles.

You can go directly to the BLAST site via ▸ http://www.ncbi.nlm.nih.gov/BLAST/. Or go the main page of NCBI (▸ http://www.ncbi.nlm.nih.gov), then select BLAST from the toolbar.

BLAST is a family of programs that allows you to input a query sequence and compare it to DNA or protein sequences in a database. The programs produce high-scoring segment pairs (HSPs) that represent local alignments between your query and database sequences.

There are four components to performing any BLAST search:

1. Selecting a sequence of interest and pasting it into the BLAST input box.
2. Selecting a BLAST program (blastp, blastn, blastx, tblastx, tblastn).
3. Selecting a database to search. A common choice is the nonredundant (nr) database, but there are many other databases.

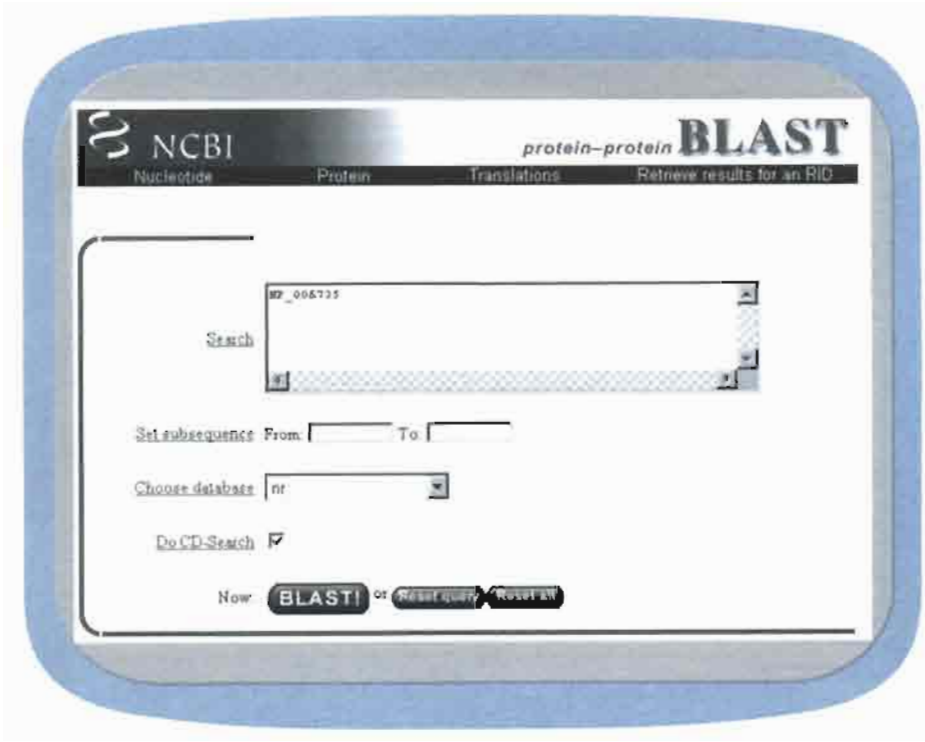


FIGURE 4.1. *Performing a protein–protein BLAST search for human RBP4.*

4. Selecting optional parameters, both for the search and for the format of the output. These options include choosing a substitution matrix, filtering of low-complexity sequences, and restricting the search to a particular set of organisms.

As we describe the steps of BLAST searching, we can begin with a specific example. Select the link "Standard protein-protein BLAST [blastp]." You will see a box marked "search"; enter the accession number for human RBP4 (NP_006735), then click the "BLAST!" button (Fig. 4.1). A new page opens called "Formatting BLAST" (Fig. 4.2). Press the "format" button (Fig. 4.2) and wait for your results. The result lists the proteins that are most closely related to RBP. We will now describe the practical aspects of BLAST searching in detail.

As of May 2003, you have searched a database of over 1.5 million protein sequences (and about 500 million amino acid residues) within several seconds.

BLAST Search Steps

Step 1: Specifying Sequence of Interest

A BLAST search begins with the selection of a DNA or protein sequence. There are three main forms of data input: (1) cutting and pasting DNA or protein sequence, (2) using sequence in the FASTA format, and (3) simply using an accession number [e.g., a RefSeq or GenBank Identification (GI) number]. A sequence in FASTA format begins with a single-line description followed by lines of sequence data. The description line is distinguished from the sequence data by a greater than (">") symbol in the first column. It is recommended that all lines of text be shorter than 80 characters in length. An example of a sequence in FASTA format was shown in Figure 2.10.

The FASTA format is further described at ▸ http://www.ncbi.nlm.nih.gov/BLAST/fasta.html. Do not confuse the FASTA format with the FASTA program, which we described briefly in Chapter 3.

For BLAST searches, your query can be in uppercase or lowercase, with or without intervening spaces or numbers. (For FASTA sequences, the format is different, and spaces in the middle are not allowed.) If the query is DNA, BLAST algorithms will search both strands.

It is often simpler and easier to input the accession number to a BLAST search. Note that the BLAST programs can recognize and ignore numbers that appear in

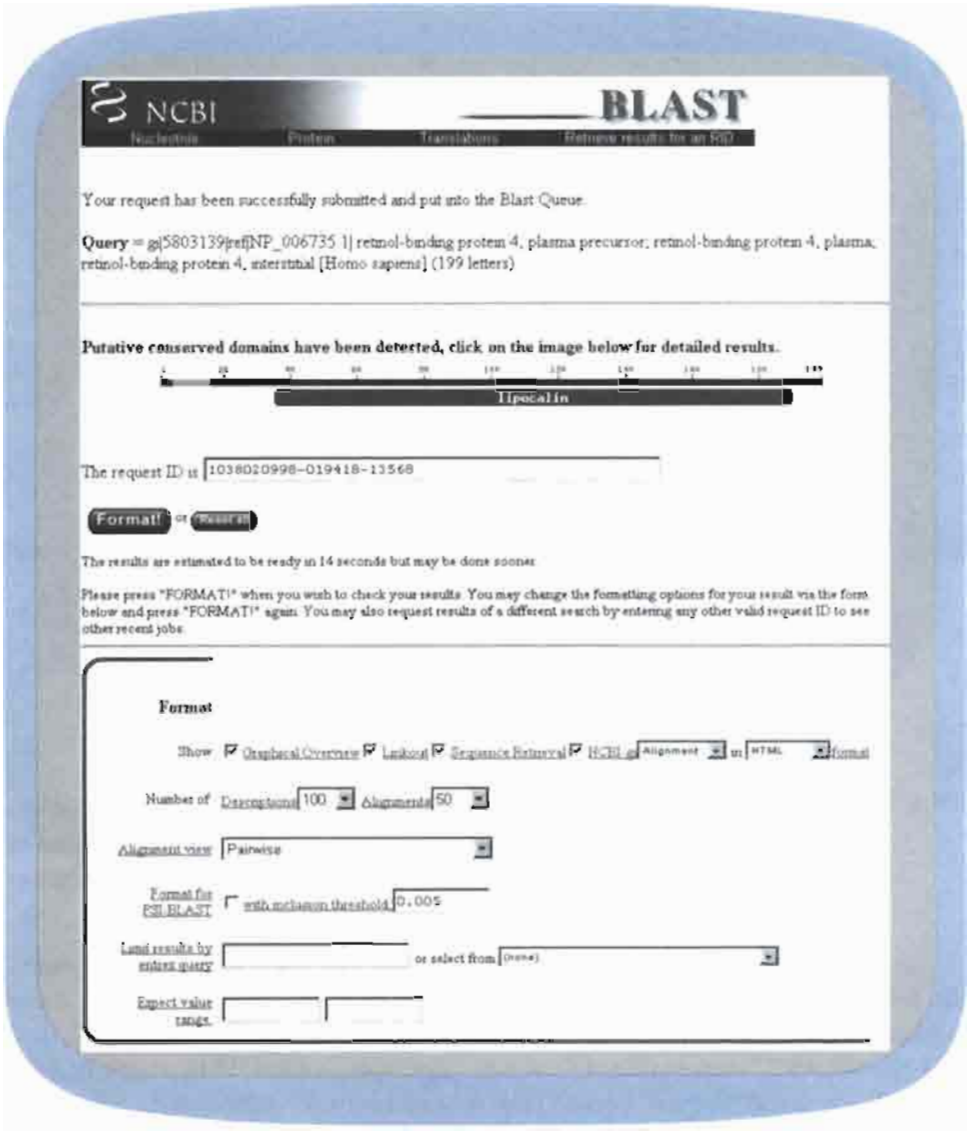


FIGURE 4.2. Formatting BLAST page.

the midst of the letters of your input sequence. The BLAST search also allows you to input a subset of an entire sequence, such as a region or domain of interest.

Step 2: Selecting BLAST Program

The NCBI BLAST family of programs includes five main programs, as summarized in Figure 4.3.

1. The first program is *blastp*. This program compares an amino acid query sequence against a protein sequence database. Note that for this type of search there are optional parameters (see below) that are specifically relevant to protein searches, such as the choice of various PAM and BLOSUM scoring matrices.
2. The second program is *blastn*. This is used to compare a nucleotide query sequence against a nucleotide sequence database.

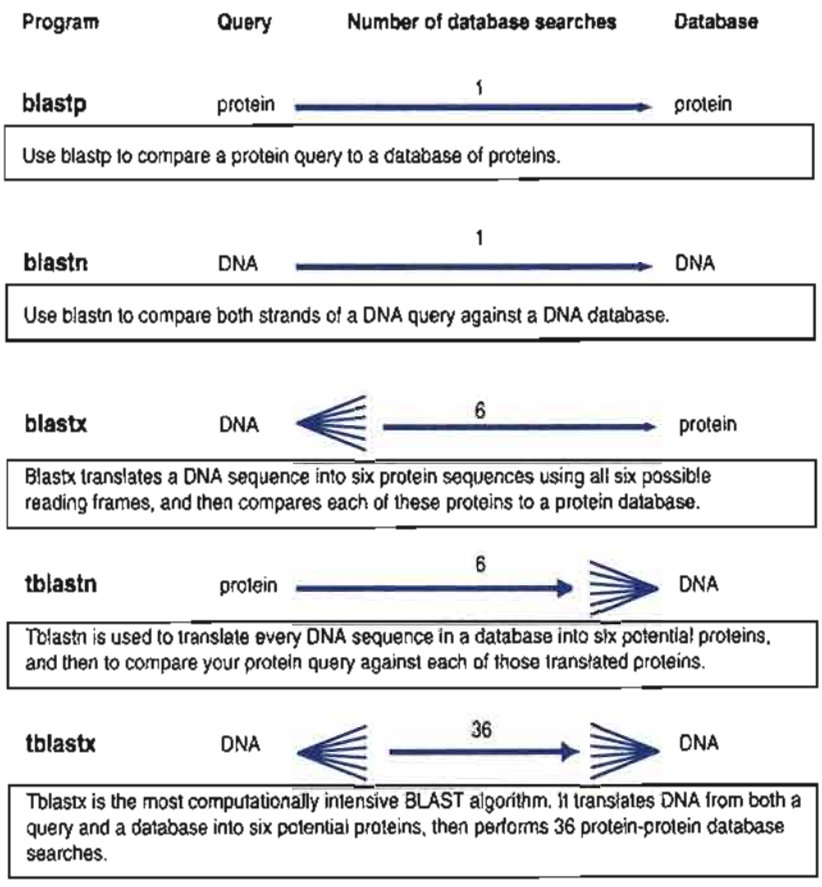


FIGURE 4.3. Overview of the five main BLAST algorithms. Note that the suffix p refers to protein (as in blastp), n refers to nucleotide, and x refers to a DNA query that is dynamically translated into six protein sequences. The prefix t refers to "translating," in which a DNA database is dynamically translated into six proteins. We will discuss the use of these BLAST algorithms later in this chapter (Fig. 4.17).

Three additional BLAST algorithms rely on the fundamental relationship of DNA to protein. Any DNA sequence can be transcribed and translated into six potential reading frames (three on the top strand and three on the bottom strand; Fig. 4.4). For BLAST searching, the query DNA sequence may be translated into potential proteins, an entire DNA database may be translated, or both. In all three cases, these algorithms perform protein–protein alignments.

3. The program *blastx* compares a nucleotide query sequence translated in all reading frames against a protein sequence database. If you have a DNA sequence and you want to know what protein (if any) it encodes, you can perform a blastx search. This automatically translates the DNA into six potential proteins (see Figs. 4.3 and 4.4). The blastx program then compares each of the six translated protein sequences to all the members of a protein database.
4. The program *tblastn* compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames. One might use this program to ask whether a DNA database encodes a protein that matches your protein query of interest. Does a query with RBP yield any matches in a database of genomic DNA from the genome-sequencing project of a particular organism?
5. The program *tblastx* compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. The tblastx program cannot be used with the main nonredundant (nr) database on the BLAST web page, because this operation is

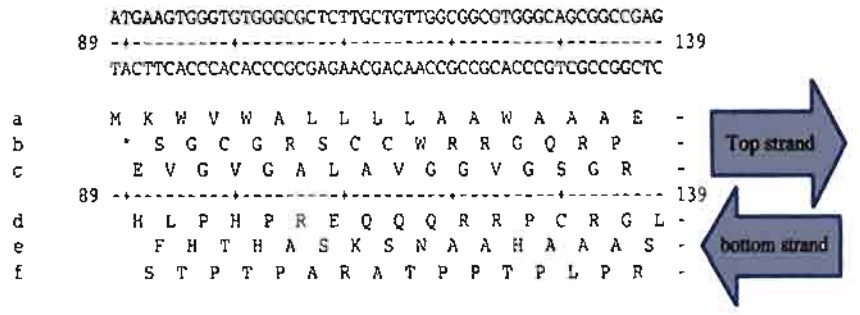


FIGURE 4.4. DNA can potentially encode six different proteins. The two strands of DNA sequence of human RBP (accession NM_006744) are shown. From the top strand, three potential proteins are encoded (frames a, b, c). The protein encoded in frame a is authentic RBP. The first codon of potential protein b, TGA, encodes a stop codon (asterisk). On the bottom strand, three additional proteins are potentially encoded (frames d, e, f). For example, the first amino acid of protein d is leucine, encoded by CTC. This figure was generated using the map program of GCG.

computationally very intensive. We will define the nr database below. Consider a situation in which you have a DNA sequence with no obvious database matches, and you want to know if it encodes a protein with even distant database matches. A blastx search would be useful to reveal such matches. But if that search fails, you might perform a tblastx search to determine whether an entire DNA database contains genes that encode proteins homologous to your query.

Step 3: Selecting a Database

Information on databases is also available at ▸http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.html.

The databases that are available for BLAST searching are listed on each BLAST page. For protein database searches (blastp and blastx), the main two options are the nr database and SwissProt. For proteins, the nr database consists of the combined protein records from GenBank, the Protein Data Bank (PDB), SwissProt, PIR, and PRF (see Chapter 2 for descriptions of these resources).

For DNA database searches (blastn, tblastn, tblastx) the options are to search the nucleotide nr database or the EST database. The nr database for DNA searches includes nucleotide sequences from GenBank, EMBL, DDBJ, and PDB. However, the nr database does not have records from the EST, STS, GSS, or high-throughput genomic sequence (HTGS) databases.

The nr databases are derived by merging several main protein or DNA databases. These databases often contain identical sequences, but only one of these sequences is retained by the nr database. (Even if two sequences in the nr database appear to be identical, they in fact have some subtle difference.) The nr databases are typically the preferred sites for searching the majority of available sequences.

A summary of all the protein databases and nucleotide sequence databases that can be searched by standard BLAST searching at NCBI is provided in Tables 4.1 and 4.2.

Step 4a: Selecting Optional Search Parameters

We will initially focus our attention on a standard protein–protein BLAST search. In addition to deciding on which sequence to input and which database to search, there are 10 more optional parameters that you can adjust (see Figs. 4.1 and 4.5).

TABLE 4-1 Protein Sequence Databases That Can Be Searched by Standard BLAST Searching

Database	Description
nr	Nonredundant GenBank coding sequences + PDB + SwissProt + PIR + PRF
Month	Sequence data released in the previous 30 days
Swissprot	Most recent release from SwissProt
Drosophila	*Drosophila* proteins from the *Drosophila* Genome Project (▸http://www.fruitfly.org)
S. cerevisiae	*Saccharomyces cerevisiae* (yeast) proteins
E. coli	*Escherichia coli* proteins
Pdb	Protein data bank at Brookhaven (▸http://www.rcsb.org/pdb/)
alu	Translations of select *Alu* repeats from REPBASE, suitable for masking *Alu* repeats from query sequences. It is available by anonymous file transfer protocol (FTP) from ncbi.nlm.nih.gov (under the/pub/jmc/alu directory).

Source: Modified from ▸http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.html#protein_databases.

1. **Do CD-Search (Fig. 4.1).** This refers to the Conserved Domain Database (CDD) that describes conserved modules or domains found in a protein. We will describe CDD in Chapter 8.

The Conserved Domain Database is at ▸http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

2. **Limit by Entrez Query (Fig. 4.5).** Any NCBI BLAST search can be limited using any terms that are used in an Entrez search. Enter the term "protease" and perform a blastp search using RBP4 as a query (NP_006735); instead of many dozens of hits, the principal database match is a lipocalin (α-1-microglobulin; NP_001624) that is fused to a protein with a serine protease inhibitor domain. BLAST searches can also be restricted by organism using a convenient pull-down menu. Some

TABLE 4-2 Nucleotide Sequence Databases That Can Be Searched by Standard BLAST Searching

Database	Description
nr	All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1, or 2 HTGS sequences); no longer "nonredundant"
Month	All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days
Drosophila genome	*Drosophila* genome provided by Celera and Berkeley *Drosophila* Genome Project (BDGP) (▸http://www.fruitfly.org/)
Dbest	Database of GenBank+EMBL+DDBJ sequences from EST divisions
Dbsts	Database of GenBank+EMBL+DDBJ sequences from STS divisions
Htgs	Unfinished high-throughput genomic sequences
Gss	Genome survey sequence, includes single-pass genomic data, exon-trapped sequences, and *Alu* PCR sequences.
S. cerevisiae	Yeast (*S. cerevisiae*) genomic nucleotide sequences
E. coli	*Escherichia coli* genomic nucleotide sequences
Pdb	Sequences derived from the three-dimensional structure from Brookhaven Protein Data Bank (▸http://www.rcsb.org/pdb/)
Vector	Vector subset of GenBank, NCBI, in ftp://ncbi.nlm.nih.gov/blast/db/
Mito	Database of mitochondrial sequences
alu	Select *Alu* repeats from REPBASE, suitable for masking *Alu* repeats from query sequences (available by anonymous FTP from ncbi.nlm.nih.gov under the /pub/jmc/alu directory)
epd	Eukaryotic Promoter Database (▸http://www.genome.ad.jp/dbget-bin/www_bfind?epd)

Source: Modified from ▸http://www.ncbi.nlm.nih.gov/blast/html/blastcgihelp.html#nucleotide_databases.

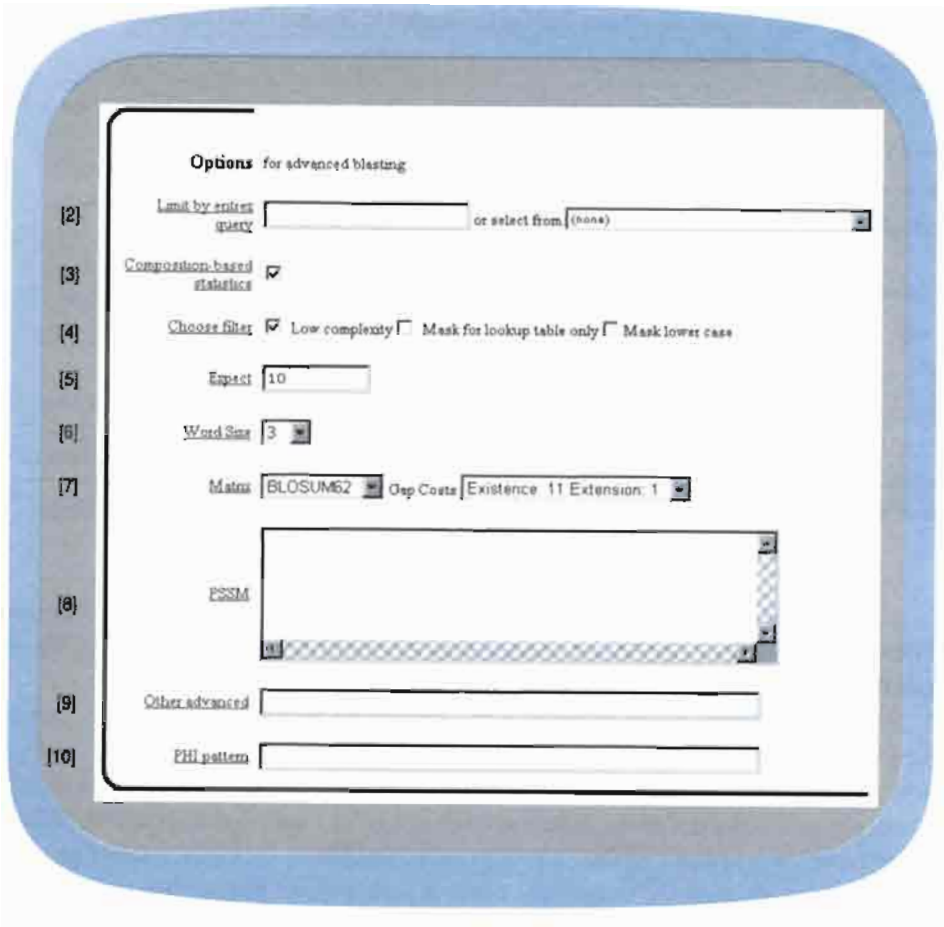


FIGURE 4.5. Options menu from a protein–protein BLAST page. Numbers in brackets refer to discussion in the text.

popular groups are Archaea, Metazoa (multicellular animals), Bacteria, Vertebrata, Eukaryota, Mammalia, Embryophyta (higher plants), Rodentia, Fungi, and Primates. BLAST searches can be restricted to any genus and species or other taxonomic grouping.

3. **Composition-Based Statistics.** This option, which is selected as default, generally improves the calculation of the *E* value statistic (see below) and increases selectivity (reduces the number of false-positive results that are returned). This option implements a slightly different scoring system for each database sequence and is applicable most importantly to the position-specific scoring matrix of PSI-BLAST (Chapter 5).

4. **Choose Filter.** Filtering masks portions of the query sequence that have low complexity (or highly biased compositions) (Wootton and Federhen, 1996). Low-complexity sequences are defined as having commonly found stretches of amino acids (or nucleotides) with limited information content. Examples are dinucleotide repeats, or regions of a protein that are extremely rich in one or two amino acids. Stretches of hydrophobic amino acid residues that form a transmembrane domain are very common, and a database search with such sequences results in many database matches that are statistically significant but biologically irrelevant. Other motifs that are masked by filtering include acidic-, basic-, and proline-rich regions. For protein sequence queries, the SEG program is used; for nucleic acid sequences, the DUST program is employed. Upon filtering a query sequence, contiguous low-complexity residues are replaced with a string of characters with the letter X (for protein sequences; Fig. 4.6) or N (for nucleic acid sequences). It may be helpful to

```
>gi|5803139|ref|NP_006735.1| retinol-binding protein 4, interstitial precursor [Homo sapiens]
 gi|132404|sp|P02753|RET4_HUMAN PLASMA RETINOL-BINDING PROTEIN PRECURSOR (PRBP) (RBP)
 gi|72085|pir||VAHU plasma retinol-binding protein precursor - human
 gi|35897|emb|CAA24959.1| (X00129) precursor RBP [Homo sapiens]
          Length = 199

 Score =  378 bits (971), Expect = e-104
 Identities = 187/199 (93%), Positives = 187/199 (93%)

Query: 1   MKWVXXXXXXXXXXXXERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAE 60
           MKWV            ERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAE
Sbjct: 1   MKWVWALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAE 60

Query: 61  FSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 120
           FSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH
Sbjct: 61  FSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDH 120

Query: 121 WIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQ 180
           WIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQ
Sbjct: 121 WIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQ 180

Query: 181 YRLIVHNGYCDGRSERNLL 199
           YRLIVHNGYCDGRSERNLL
Sbjct: 181 YRLIVHNGYCDGRSERNLL 199
```

FIGURE 4.6. Output of a BLAST search with the filter option selected. Note that 12 amino acids in the query sequence have been replaced with the character X. These residues have been filtered because they have low complexity. In this case, the 12 amino acids form part of a signal sequence and are hydrophobic; in the absence of filtering, many spurious hundreds of database matches would be created to other proteins that also have stretches of hydrophobic amino acids. For nucleic acid BLAST queries, low-complexity sequences (such as strings of individual nucleotides or dinucleotides) are replaced by the letter N.

turn off the filtering option when using a very short DNA or protein sequence as a query. Note that query sequences are filtered, but not databases.

Adjusting the filtering option can have dramatic effects on BLAST search results. When a human proline-rich protein (NP_036522) is searched using blastp nr, there are eight database matches of which two appear significant (Fig. 4.7). When

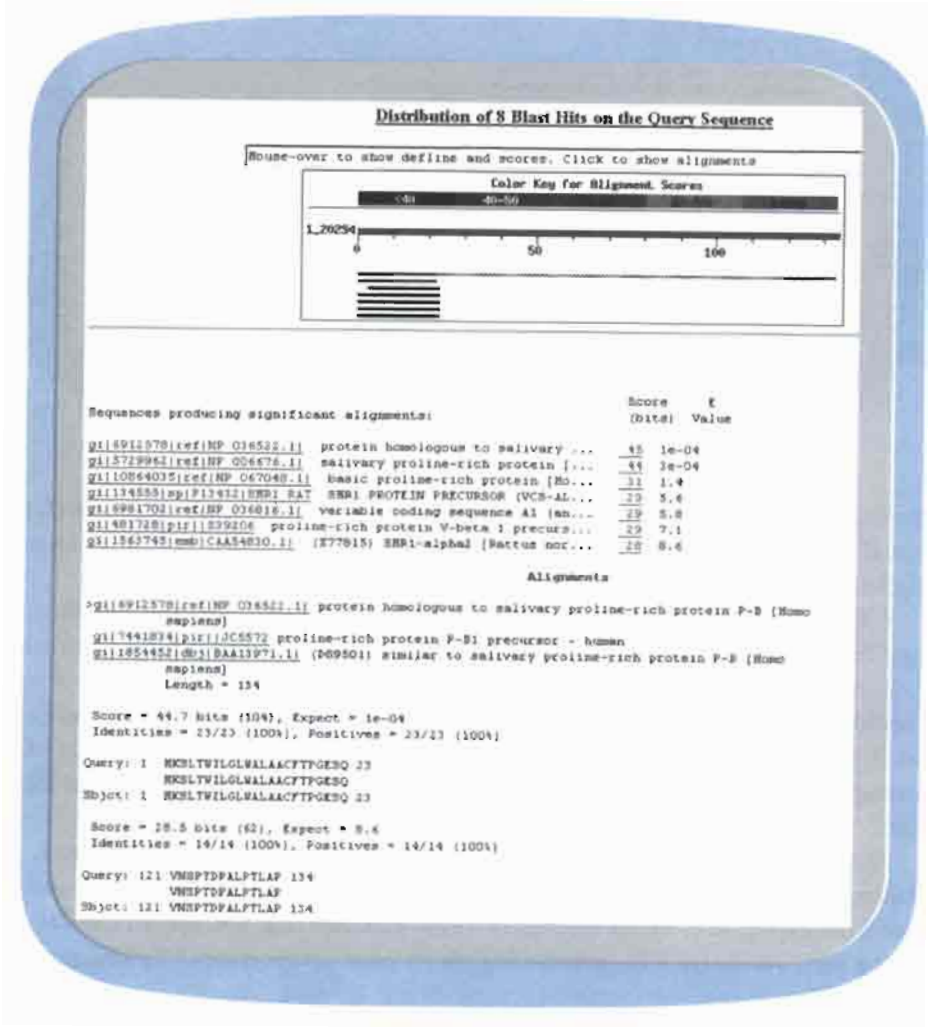


FIGURE 4.7. Result of a blastp search of the nr database using a human proline-rich salivary protein (NP_036522) as a query. Note that there are only eight database hits, including two that are closely related to the query. Also, the majority of the protein (from amino acids 24–120) has no database hits. This is because the central region of the protein has been filtered by default.

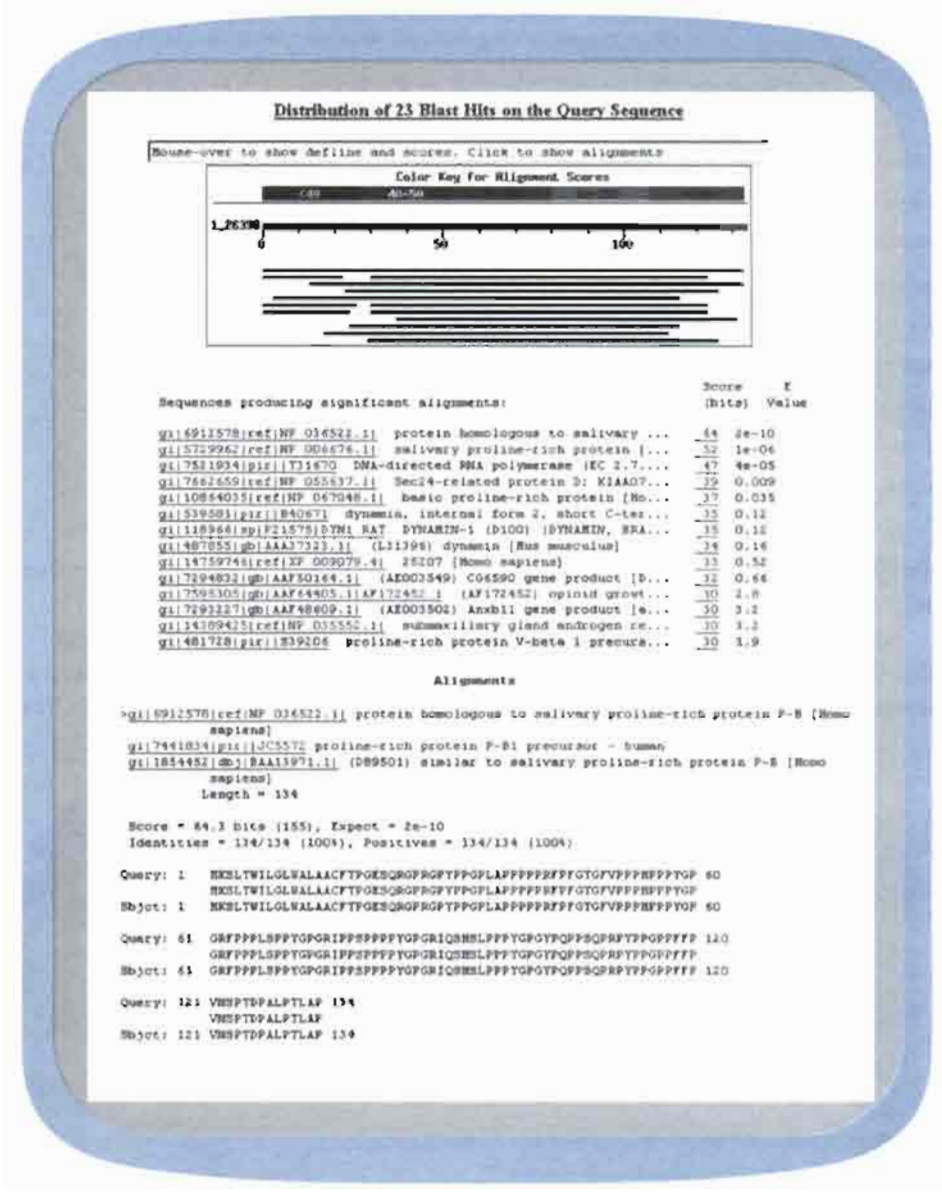


FIGURE 4.8. Results of a blastp search with a proline-rich protein in which the filtering is turned off. Note that there are more database returns with matches across the length of the protein (compare to Fig. 4.7). The first alignment shows the extremely proline-rich nature of the protein, which is masked by the default BLAST filter.

filtering is turned off, there are now 23 matches, including other proline-rich proteins (Fig. 4.8). In general, filtering is useful to avoid receiving spurious database matches, but in some cases authentic matches may be missed.

5. **Expect.** The expect value E is the number of different alignments with scores equal to or greater than some score S that are expected to occur in a database search by chance. Look at the best match in Fig. 4.7. The score is 45 bits, and the E value is 1e-04 (that is, 1×10^{-4}, or 0.0001). This indicates that based on the particular search parameters used (including the size of the database and the choice of the scoring matrix), a score of 45 or better is expected to occur by chance 1 in 10,000 times.

The expect value is sometimes also referred to as the expectation value.

The default setting for the expect value is 10 for blastn, blastp, blastx, and tblastn. At this E value, 10 hits with scores equal to or better than the alignment score S are expected to occur by chance. (This assumes that you search the database using a random query with similar length to your actual query.) By changing the

expect option to a lower number, fewer database hits are returned; fewer chance matches are reported. Increasing E returns more hits. We will describe the E value in more detail in a discussion of BLAST search statistics later in this chapter.

6. **Word Size.** For protein searches, a window size of 3 (default) or 2 may be set. When a query is used to search a database, the BLAST algorithm first divides the query into a series of smaller sequences (words) of a particular length (word size), as will be described below. For blastp, a larger word size yields a more accurate search. For any word size, matches made to each word are then extended to produce the BLAST output. In practice, there is rarely a need to change the word size for protein searches.

For nucleotide searches, the default word size is 11 and can be raised (word size 15) or reduced (word size 7). Lowering the word size yields a more accurate but slower search.

7. **Matrix.** Five amino acid substitution matrices are available for blastp protein–protein searches: PAM30, PAM70, BLOSUM45, BLOSUM62 (default), and BLOSUM80. Some alternative BLAST servers (discussed in Chapter 5, on advanced BLAST searching) offer many more choices for substitution matrices such as PAM250. It is usually advisable to try a BLAST search using several different scoring matrices. For example, as described in Chapter 3, PAM40 and PAM250 matrices (Fig. 3.16) have entirely distinct properties as scoring matrices for sequences sharing varying degrees of similarity.

8–10. The position-specific scoring matrix (PSSM), other advanced options, and the pattern hit-initiated (PHI) matrix will be described in Chapter 5.

Try doing a blastp search with RBP (NP_006735) using E values of 1000, 10 (the default value), or 0.0001. In each case the BLAST algorithm identifies the same number of database matches, whether they are significant or not. But the blastp program reports in its output a list of about 750 sequences with an E value better than 1000, 140 sequences with an E value better than 10, and 50 sequences with an E value better than 0.0001. We will see later in Figure 4.16 how you can find this information at the very bottom of any BLAST search output page.

Step 4b: Selecting Optional Formatting Parameters

There are many options for formatting the output of a BLAST search. These are illustrated by performing a protein–protein BLAST search with human RBP4 (NP_006735) and restricting the organism to the rat (*Rattus norvegicus*). The results of the search occur in several main parts. In the top (Fig. 4.9), details of the search are provided, including the type of BLAST search, a description of the query and the database that were searched, and a taxonomy link to the results organized by species.

The middle portion of a typical BLAST output provides a list of database sequences that match the query sequence (Fig. 4.10). A graphical overview provides a color-coded summary. Each bar drawn below the map represents a database protein or nucleic acid sequence that matches the query sequence. The position of each bar relative to the linear map of the query allows the user to see instantly the extent to which the database matches align with a single or multiple regions of the query. The most similar hits are shown at the top in red. Hatched areas (when present) correspond to the nonsimilar sequence between two or more distinct regions of similarity found within the same database entry.

The alignments are described by a list of one-line summaries called "descriptions." The description lines are sorted by increasing E value; thus the most signficant alignments (lowest E values) are at the top. The description consists of four columns (from the left):

1. Identifier for the database sequence, the following being some of the common sequence identifiers:

   RefSeq (e.g., ref|NP_006735.1|),

   Protein database (e.g., pdb|1RBP|),

FIGURE 4.9. *Top portion of a BLAST output describes the search that was performed (BLASTP 2.2.1 in this case), the query (human RBP), and the database that was searched (nr and its size). A link is provided to a taxonomy report that organizes the search results by species.*

Protein Information Resource (pir||A27786),

Swiss-Prot database (sp|P27485|),

GenBank (e.g., gb|AAF69622.1|),

European Molecular Biology Lab (EMBL) (e.g., emb|CAB64947.1|), and

Database of Japan (DBJ)(dbj|BAA13453.1|);

2. Brief description of the sequence
3. The bit score of the highest scoring match found for each database sequence (bit scores will be defined below)
4. The expect value E. The identifer is linked to the full GenBank entry. Clicking on the score in a given description line will take the user to the corresponding sequence alignment. The alignment can also be reached by scrolling down the output page.

The lower portion of a BLAST search output consists of a series of pairwise sequence alignments, such as the one in Figure 4.6. Here, one can inspect the match between the query (input sequence) and the subject (i.e., the particular database match that is aligned to the query). Four scoring measures are provided: the bit score, the expect score, the percent identity, and the positives (percent similarity).

Without reperforming an entire BLAST search, the format options can be modified to provide a range of different output options (Fig. 4.11). The number of descriptions and of alignments can be modified. Also, there are several options for visualizing the aligned sequences as a multiple sequence alignment (Fig. 4.12). This is an especially useful way to identify specific amino acid residues that are conserved (or divergent) within a protein or DNA family.

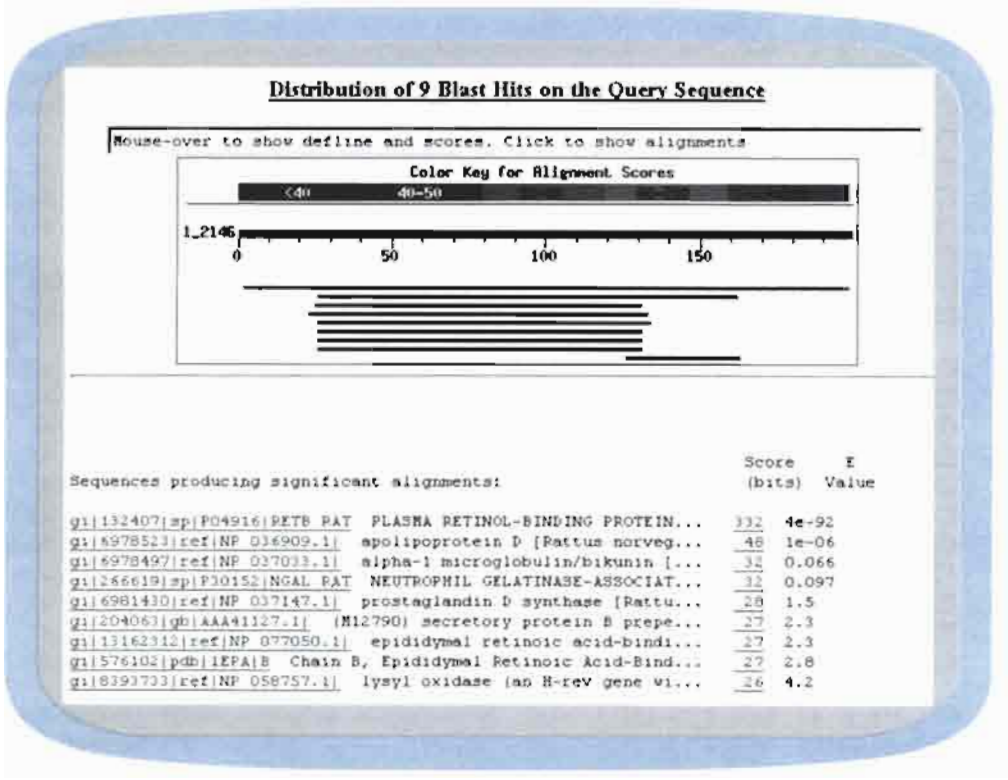


FIGURE 4.10. Middle portion of a typical BLAST output provides a graphical output of the results and then a list of those database sequences. In the graphical overview, the database matches are color coded to indicate relatedness (based on alignment score), and the length of each line corresponds to the region in which that sequence aligns with the query sequence. The list of sequences that align to the query sequence includes links to that database entry and to the alignment. The bit score and E value for each alignment are also provided. Note that the best matches at the top of the list have large bit scores and small E values.

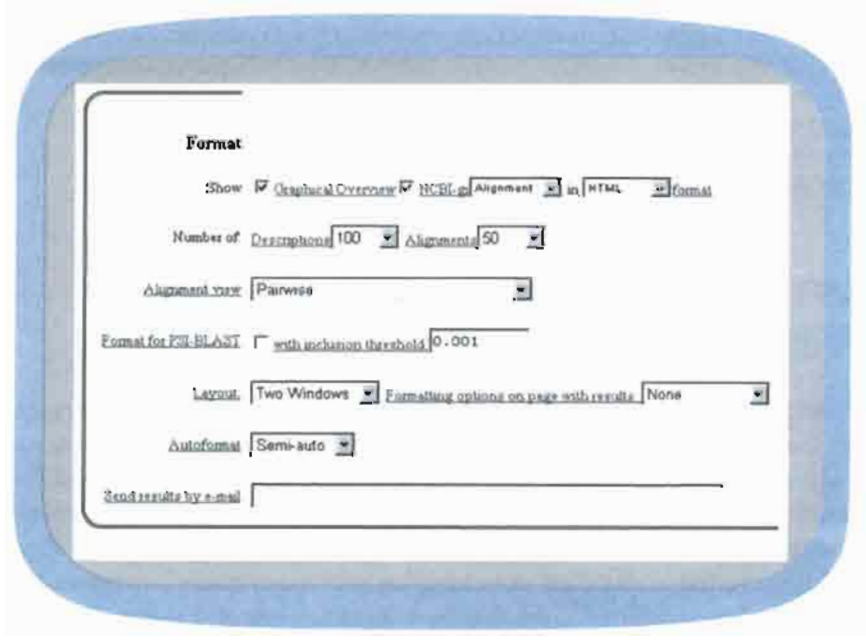


FIGURE 4.11. Format options for a protein–protein BLAST search.

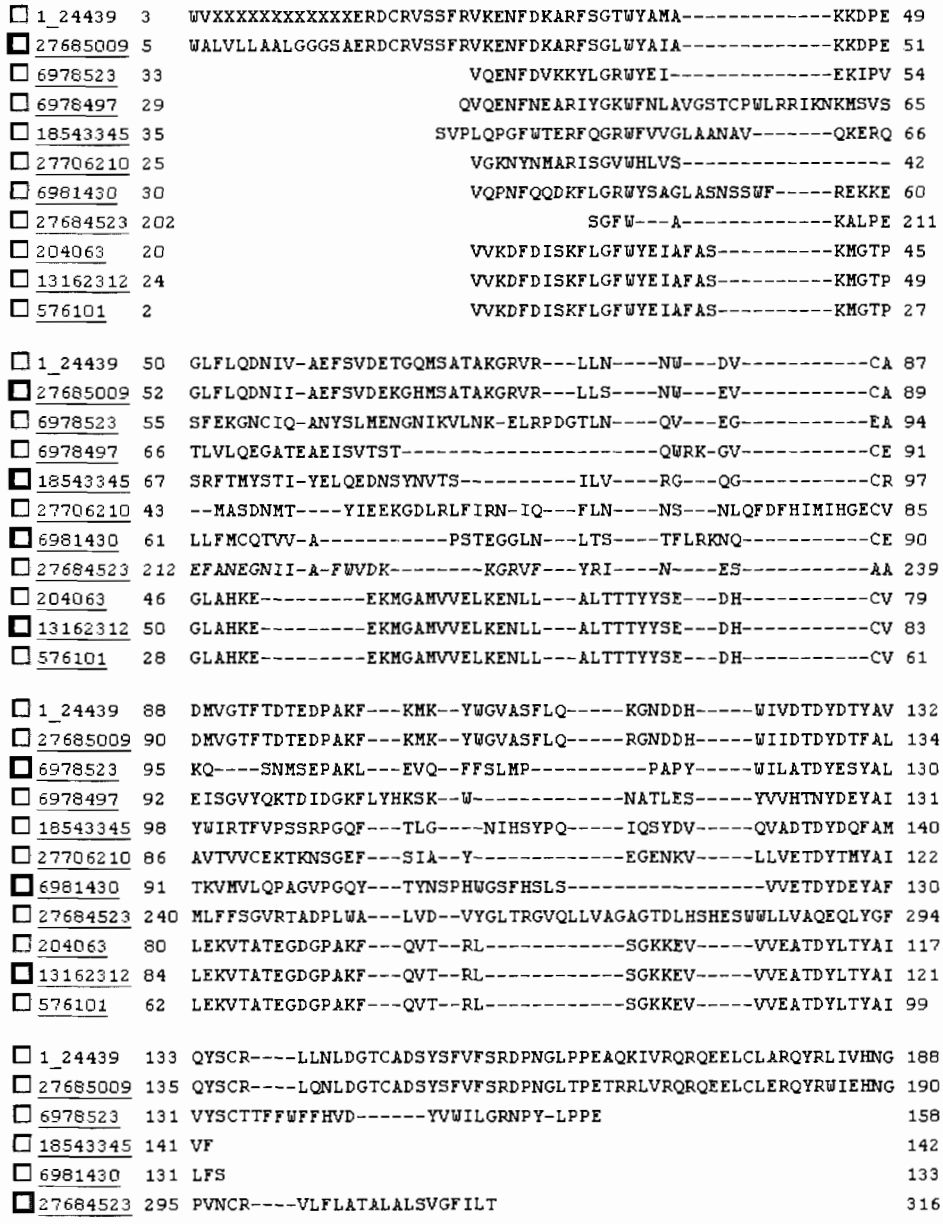


FIGURE 4.12. *Output of a BLAST search using the format option "flat query-anchored without identities." This provides a multiple sequence alignment of the BLAST search results. Other output format options are available, allowing the user to inspect regions of similarity as well as divergent regions within protein families.*

BLAST Algorithm Uses Local Alignment Search Strategy

The BLAST search identifies the matches in a database to an input query sequence. Global similarity algorithms optimize the overall alignment of two sequences. These algorithms, including the GAP program (Chapter 3), are best suited for finding matches consisting of long stretches of low similarity. In contrast, local similarity algorithms such as BLAST identify relatively short alignments. Local alignment is a useful approach to database searching because many query sequences have domains, active sites, or other motifs that have local but not global regions of similarity to other proteins. Also, databases typically have fragments of DNA and protein sequences that can be locally aligned to a query.

BLAST Algorithm Parts: List, Scan, Extend

The BLAST search algorithm finds a match between a query and a database sequence and then extends the match in either direction (Altschul et al., 1990, 1997). The search results consist of both highly related sequences from the database as well as marginally related sequences, along with a scoring scheme to describe the degree of relatedness between the query and each database hit. This algorithm can be described in three phases (Fig. 4.13):

1. BLAST compiles a preliminary list of pairwise alignments, called word pairs.
2. The algorithm scans a database for word pairs that meet some threshold score T.
3. BLAST extends the word pairs to find those that surpass a cutoff score S, at which point those hits will be reported to the user. Scores are calculated from scoring matrices (such as BLOSUM62) along with gap penalties.

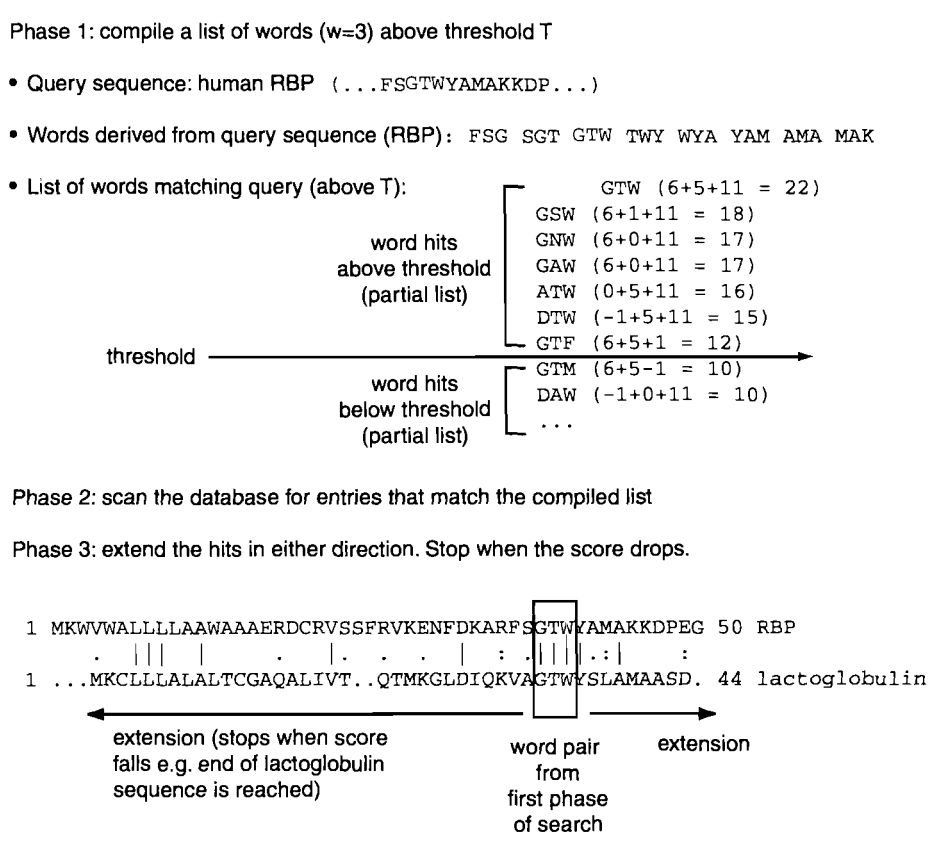


FIGURE 4.13. *Schematic of the original BLAST algorithm. In the first phase a query sequence (such as human RBP4) is analyzed with a given word size (e.g., $w = 3$), and a list of words is compiled having a threshold score (e.g., $T = 11$). Several possible words are listed in the figure (from FSG to MAK); in an actual BLAST search there may be hundreds or thousands of words compiled. For a given word, such as the portion of the query sequence consisting of GTW, a list of words is compiled with scores greater than or equal to some threshold T (e.g., 11). In this example, nine words are shown along with their scores from a BLOSUM62 matrix; seven of these are above the threshold, and two are below. In phase 2, a database is scanned to find entries that match the compiled word list. In phase 3, the database hits are extended in both directions to obtain a high-scoring segment pair (HSP). If the HSP score exceeds a particular cutoff score S, it is reported in the BLAST output. Note that in this particular example the word pair that triggers the extension step happens to be an exact match (see boxed residues GTW). However, the main idea of the threshold T is to also allow nonexact word hits to trigger an extension.*

In the BLAST papers by Steven Altschul, David Lipman, and colleagues, the threshold parameter is denoted T (Altschul et al, 1990, 1997). In the BLAST program, the threshold parameter is called f.

The two-hit approach greatly speeds up the time required to do a BLAST search. Compared to the one-hit approach, the two-hit method generates on average about three times as many hits, but the algorithm then needs to perform only one-seventh as many extensions (Altschul et al., 1997). For the parameter A, the default value is 40.

You can try this by using the advanced option command "-f 16" (or choose other f values). Alternatively try it with NetBLAST, which is a version of BLAST you can download from the NCBI website onto a variety of platforms and use via FTP. Find it at ▸ ftp://ncbi.nlm.nih.gov/blast/network/netblast. Once you install it, create a file with your favorite query sequence (e.g., rbp4.fasta) and save it on your hard drive. Then perform the command "$ blastcl3 -p blastp -d nr -i rbp4.fasta-o output1 -f 11." This command specifies the program (blastp), the database you would like to search (nr), the input query file (rbp4.fasta), the output file (here called output1), and finally the threshold value f. Your search will be sent to NCBI by FTP, and the result will be returned promptly.

Note that for DNA database searches with BLAST, threshold scores are not used in association with words. Instead, the algorithm demands exact word matches. Lowering the word length effectively lowers the threshold score. Thus, lowering the word length induces a slower, more accurate search.

In the first phase, the BLAST algorithm compiles a list of "words" of a fixed length w that are derived from the query sequence. A threshold value T is established for the score of aligned words. Those words either at or above the threshold are defined as hits; those words below threshold are not further pursued. For protein searches the word size typically has a default value of 3, while for DNA the default word size is 11. The word size parameter can be modified by the BLAST user, as described above (see option 6). The threshold score T can be lowered to identify more initial pairwise alignments. This will increase the time required to perform the search and may increase the sensitivity.

In the second phase, after compiling a list of word pairs at or above threshold T, the BLAST algorithm scans a database for hits. This requires BLAST to find database entries that correspond to words on the compiled list. In the original implementation of BLAST, one hit was sufficient. In the current versions of BLAST, the algorithm seeks two separate word pairs (i.e., two nonoverlapping hits) within a certain distance A from each other. It then generates an ungapped extension of these hits (Altschul et al., 1997).

In the third phase, BLAST extends hits to find alignments called high-scoring segment pairs (HSPs). For sufficiently high-scoring alignments, a gapped extension is triggered. The extension process is terminated when a score falls below a cutoff.

In summary, the main strategy of the BLAST algorithm is to compare a protein or DNA query sequence to each database entry and to form pairwise alignments (HSPs). As a heuristic algorithm, BLAST is designed to offer both speed and sensitivity. When the threshold parameter is raised, the speed of the search is increased, but fewer hits are registered, and so distantly related database matches may be missed. When the threshold parameter is lowered, the search proceeds far more slowly, but many more word hits are evaluated, and thus sensitivity is increased.

We can demonstrate the effect of different threshold levels on a blastp search by changing the f parameter from its default value (11) to either a very low value (5) or a high one (16). The results are dramatic (Table 4.3). With the default threshold value of 11, there are about 130 million hits to the database and 5 million extensions. When the threshold is 5, there are about 2.2 billion hits to the database and 590 million extensions. With the threshold raised to 17, there are only 12 million hits and 61,000 extensions. The number of successful extensions is also greater with the low threshold value. The final results of the search are not dramatically different with the default value compared to the lowered threshold, as the number of gapped HSPs is comparable. With the high threshold some matches were missed. This supports the conclusion that a lower threshold parameter yields a more accurate search,

TABLE 4-3 Effect of Changing Threshold Values on Blastp Search

These three searches were done using the NCBI blastcl3 program (NetBLAST) using *RBP4* (NP_006735) as a query. When the threshold parameter (*f*) was changed to 5, there were over 2 billion hits and 589 million extensions, ultimately producing more gapped HSPs (146) than found with the higher threshold value of 17

	$f = 11$ (default)	$f = 5$	$f = 17$
Number of sequences in database	1,046,476	1,046,476	1,046,476
Number of hits to database	129,839,417	2,200,945,350	12,002,487
Number of extensions	5,198,652	589,935,555	61,838
Number of successful extensions	8,377	13,145	1,117
Number of HSPs gapped	145	146	93

though a slower one. This trade-off between sensitivity and speed is central to the BLAST algorithm.

BLAST Algorithm: Local Alignment Search Statistics and *E* Value

We care about the statistical significance of a BLAST search because we want some quantitative measure of whether the alignments represent significant matches or whether they would be expected to occur by chance alone. The BLAST programs provide such statistics (Altschul and Gish, 1996).

For global alignments, we described a scramble test (Chapter 3) in which one sequence is scrambled many times and a Z value (number of standard deviations from the mean) for the alignment score is generated. However, the distribution of globally aligned sequences is poorly understood, and the shape of the distribution of scores is not known. In particular, the scores may not be randomly distributed. Thus, the statistical analysis of global alignments is poorly understood, and it is not valid to describe the probability that an alignment score is significantly different than would be expected by chance.

For local alignments (including BLAST searches), more rigorous statistical tests have been developed (Altschul and Gish, 1996; Altschul et al., 1990, 1994, 1997; Pagni and Jongeneel, 2001). We have described how local, ungapped alignments between two protein sequences are analyzed as HSPs. Using a substitution matrix, specific probabilities are assigned for each aligned pair of residues, and a score is obtained for the overall alignment. For the comparison of a query sequence to a database of random sequences of uniform length, the scores can be plotted and shown to have the shape of an extreme value distribution (see Fig. 4.14, where it is compared to the normal distribution). The normal, or Gaussian, distribution forms the familiar, symmetric bell-shaped curve. The extreme value distribution is skewed to the right, with a tail that decays in x (rather than x^2, which describes the decay of the normal distribution). The properties of this distribution are central to our understanding of BLAST statistics because they allow us to evaluate the likelihood that the highest scores from a search (i.e., the values at the right-hand tail of the distribution) occurred by chance. A search of RBP4 against a protein database using a

Because of the rapid tailing of the normal distribution in x^2, if we tried to use the normal distribution to describe the significance of a BLAST search result—for example by estimating how many standard deviations above the mean a search result occurs—we would tend to overestimate the significance of the alignment.

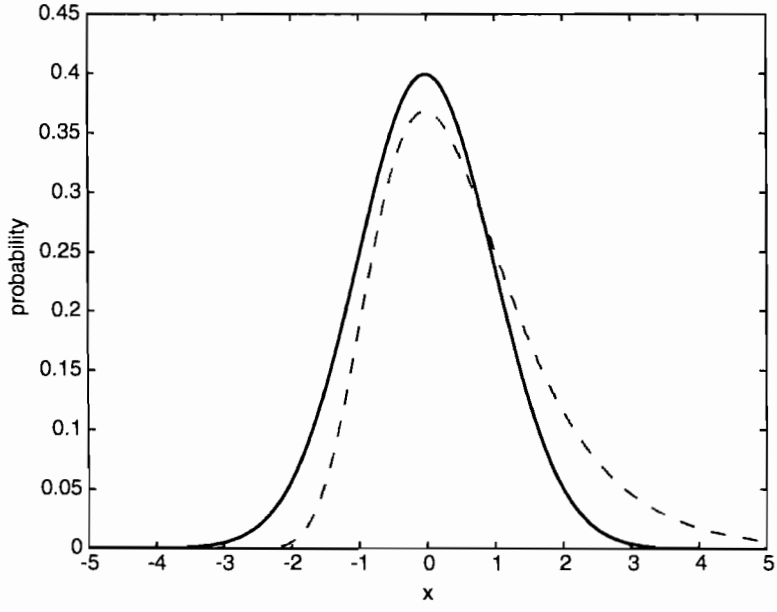


FIGURE 4.14. *Normal distribution (solid line) is compared to the extreme value distribution (dotted line). Comparing a query sequence to a set of uniform-length random sequences usually generates scores that fit an extreme-value distribution (rather than a normal distribution). The area under each curve is 1. For the normal distribution, the mean (μ) is centered at zero, and the probability Z of obtaining some score x is given in terms of units of standard deviation (σ) from x to the mean: $Z = (x - \mu)/\sigma$. In contrast to the normal distribution, the extreme value distribution is asymmetric, with a skew to the right. It is fit to the equation $f(x) = (e^{-x})(e^{-e^{-x}})$. The shape of the extreme value distribution is determined by the characteristic value u and the decay constant $\lambda (u = 0, \lambda = 1)$.*

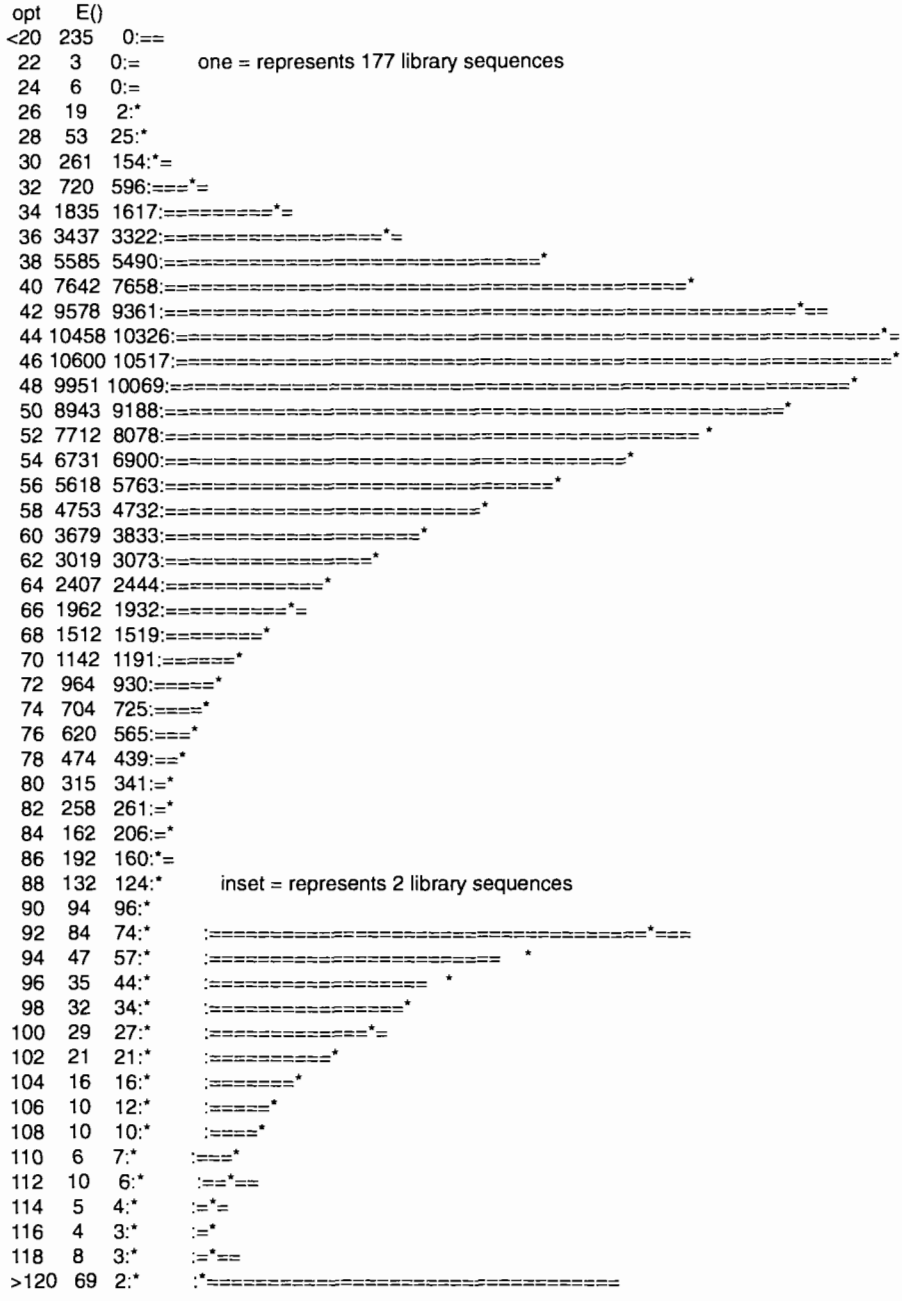


FIGURE 4.15. *Distribution of scores from a search of a query* (RBP4) *against a protein database assumes an extreme-value distribution. This search was performed using the SSEARCH implementation of the Smith and Waterman local alignment algorithm at the DNA Database of Japan web server (▸http://www.ddbj.nig.ac.jp/). The distribution of normalized similarity scores (indicated with = symbols) fits the extreme-value distribution (indicated with * symbols).*

web-based implementation of the Smith–Waterman algorithm shows an extreme-value distribution (Fig. 4.15).

We will next examine the extreme value distribution so that we can derive a formula (Equation 4.5 below) that describes the likelihood that a particular BLAST score occurs by chance.

The characteristic value u relates to the maximum of the distribution, although it is not the mean μ (mu).

The shape of the extreme value distribution shown in Figures 4.14 and 4.15 is described by two parameters: the characteristic value u and the decay constant λ. The extreme value distribution is sometimes called the Gumbel distribution, after the person who described it in 1958. The application of the extreme value distribution to BLAST searching has been reviewed by Altschul and colleagues (1994, 1996) and others (Pagni and Jongeneel, 2001). For two random sequences m

and n, the cumulative distribution function of scores S is described by the formula

$$P(S < x) = \exp(-e^{-\lambda(x-u)}) \tag{4.1}$$

To use this equation, we need to know (or estimate) the values of the parameters u and λ. For ungapped alignments, the parameter u is dependent on the lengths of the sequences being compared and is defined as

$$u = \frac{\ln Kmn}{\lambda} \tag{4.2}$$

In Equation 4.2, m and n refer to the lengths of the sequences being compared and K is a constant. Combining Equations 4.1 and 4.2, the probability of observing a score equal to or greater than x by chance is given by the formula

$$P(S \geq x) = 1 - \exp(-Kmne^{-\lambda x}) \tag{4.3}$$

Our goal is to understand the likelihood that a BLAST search of an entire database produces a result by chance alone. The number of ungapped alignments with a score of at least x is described by the parameter $Kmne^{-\lambda x}$. In the context of a database search, m and n refer to the length (in residues) of the query sequence and the length of the entire database, respectively. The product $m \cdot n$ defines the size of the search space. The search space represents all the sites at which a query sequence can be aligned to any sequence in the database. Because the ends of a sequence are not as likely to participate in an average-sized alignment, the BLAST algorithm calculates the effective search space in which the average length of an alignment L is subtracted from m and n (Altschul and Gish, 1996):

$$\text{Effective search space} = (m - L) \cdot (n - L) \tag{4.4}$$

We will see how BLAST uses this definition of search space in Figure 4.16 below.

We now arrive at the main mathematical description of the significance of scores from a BLAST search. The expected number of HSPs having some score S (or better) by chance alone is described using the equation

$$E = Kmne^{-\lambda S} \tag{4.5}$$

Here, E refers to the expect value, which is the number of different alignments with scores equivalent to or better than S that are expected to occur by chance in a database search. This provides an estimate of the number of false-positive results from a BLAST search. From Equation 4.5 we see that the E value depends on the score and λ, which is a parameter that scales the scoring system. Also, E depends on the length of the query sequence and the length of the database. The parameter K is a scaling factor for the search space. The parameters K and λ were described by Karlin and Altschul (1990) and so are often called Karlin-Altschul statistics.

Equation 4.5 is described online at ▸http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html#head2.

Note several important properties of Equation 4.5:

- The value of E decreases exponentially with increasing S. The score S reflects the similarity of each pairwise comparison and is based in part upon the scoring matrix selected. Higher S values correspond to better alignments; we saw in Figures 4.7, 4.8, and 4.10 that BLAST results are ranked by score. Thus, a high score on a BLAST search corresponds to a low E value. As E approaches zero, the probability that the alignment occurred by chance approaches zero. We will relate the E value to probability (P) values below.

- The expected score for aligning a random pair of amino acids must be negative. Otherwise, very long alignments of two sequences could accumulate large positive scores and appear to be significantly related when they are not.
- The size of the database that is searched–as well as the size of the query–influences the likelihood that particular alignments will occur by chance. Consider a BLAST result with an E value of 1. This value indicates that in a database of this particular size one match with a similar score is expected to occur by chance. If the database were twice as big, there would be twice the likelihood of finding a score equal to or greater than S by chance.
- The theory underlying Equation 4.5 was developed for ungapped alignments. However, it can be successfully applied to gapped local alignments as well (such as the results of a BLAST search). For gapped alignments λ and K cannot be calculated analytically (as they are for ungapped alignments), but instead they must be estimated by simulation.

Making Sense of Raw Scores with Bit Scores

Bit scores are displayed in a column to the right in the BLAST output window, next to corresponding E values for each database match.

A typical BLAST output reports both E values and scores. There are two kinds of scores: raw and bit scores. Raw scores are calculated from the substitution matrix and gap penalty parameters that are chosen. The bit score S' is calculated from the raw score by normalizing with the statistical variables that define a given scoring system. Therefore, bit scores from different alignments, even those employing different scoring matrices in separate BLAST searches, can be compared. A raw score from a BLAST search must be normalized to parameters such as the size of the database being queried. The raw score is related to the bit score by the equation

$$S' = \frac{\lambda S - \ln K}{\ln 2} \tag{4.6}$$

where S' is the bit score, which has a standard set of units. The E value corresponding to a given bit score is given by

$$E = mn \times 2^{-S'} \tag{4.7}$$

Why are bit scores useful? First, raw scores are unitless and have little meaning alone. Bit scores account for the scoring system that was used and describe the information content inherent in a pairwise alignment. Thus they allow scores to be compared between different database searches, even if different scoring matrices are employed. Second, bit scores can tell you the E value if you know the size of the search space, $m \cdot n$. (The BLAST algorithms use the effective search space size, described above.)

BLAST Algorithm: Relation between *E* and *P* Values

The P value is the probability of a chance alignment occurring with the score in question or better. It is calculated by relating the observed alignment score S to the expected distribution of HSP scores from comparisons of random sequences of the same length and composition as the query to the database. The most highly significant P values are those close to zero. The P and E values are different ways of representing the significance of the alignment. The probability of finding an HSP with a given E value is

$$P = 1 - e^{-E} \tag{4.8}$$

Table 4.4 lists several P values corresponding to E values. While BLAST reports E values rather than P values, the two measures are nearly identical, especially for very small values associated with strong database matches. An advantage of using E values is that it is easier to think about E values of 5 versus 10 rather than 0.99326205 versus 0.99995460.

A P value below 0.05 is traditionally used to define statistical significance (i.e., to reject the null hypothesis that your query sequence is not related to any database sequence). If the null hypothesis is true, then 5% of all random alignments will result in an apparently significant score. Thus an E value of 0.05 or less may be considered significant.

It is also possible to approach E values with conservative corrections. When you search a database with 1 million proteins, it is likely that some high-scoring alignments will occur by chance. While the E value accounts for the statistical significance of local alignments, some researchers consider it appropriate to adjust the significance level α from 0.05 to some lower value. One approach, suggested by NCBI, is to define a normalized score S' as statistically significant if it exceeds log N/E, where N is the size of the search space and E is the expect value.

Another approach is to correct α by dividing by the number of database comparisons, such as 1 million (thus $\alpha = 0.05/10^6$, or 5×10^{-8}). This is a conservative correction that accounts for the likelihood of finding strong similarities by chance in a comparison of a query to very large numbers of database sequences. In analyses of completed microbial genomes, BLAST or FASTA search E values were reported as significant if they were below 10^{-4} (Ferretti et al., 2001) or below 10^{-5} (Chambaud et al., 2001; Tettelin et al., 2001; Ermolaeva et al., 2001). In the public consortium analysis of the human genome, Smith-Waterman alignments were reported with an E value threshold of 10^{-3}, and tblastn searches used a threshold of 10^{-6} (International Human Genome Sequencing Consortium, 2001). These various values reflect different adjustments to the P value. We will encounter the idea of correcting the significance levels based on multiple tests in Chapters 6 and 7 (gene expression and microarray data analysis).

TABLE 4-4 Relationship of *E* to *P* Values in BLAST Using Equation 4.8

Small E values (0.05 or less) correspond closely to the P values

E	P
10	0.99995460
5	0.99326205
2	0.86466472
1	0.63212056
0.1	0.09516258
0.05	0.04877058
0.001	0.00099950
0.0001	0.0001000

Some BLAST servers (such as those at the European Molecular Biology Laboratory; see Chapter 5) use P values in the output.

See ▸http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/rules.html for a description of significant scores.

Gapped Alignments in BLAST

A gap is a space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another (Chapter 3). Since a single mutational event may cause the insertion or deletion of more than one residue, the presence of a gap is frequently ascribed more significance than the length of the gap. Hence the gap introduction is penalized heavily, whereas a lesser penalty is ascribed to each subsequent residue in the gap. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment.

Gap scores are typically calculated as the sum of G, the gap-opening penalty, and L, the gap extension penalty. For a gap of length n, the gap cost would be $G + Ln$. The choice of gap costs is typically 10–15 for G and 1–2 for L. These are called affine gap penalties, in which the penalty for introducing a gap is far greater than the penalty for extending one.

Getting to Bottom of BLAST Search

Each of the various BLAST algorithms (including BLAST 2 Sequences) provides a summary of the search statistics at the bottom of the output page. An example in

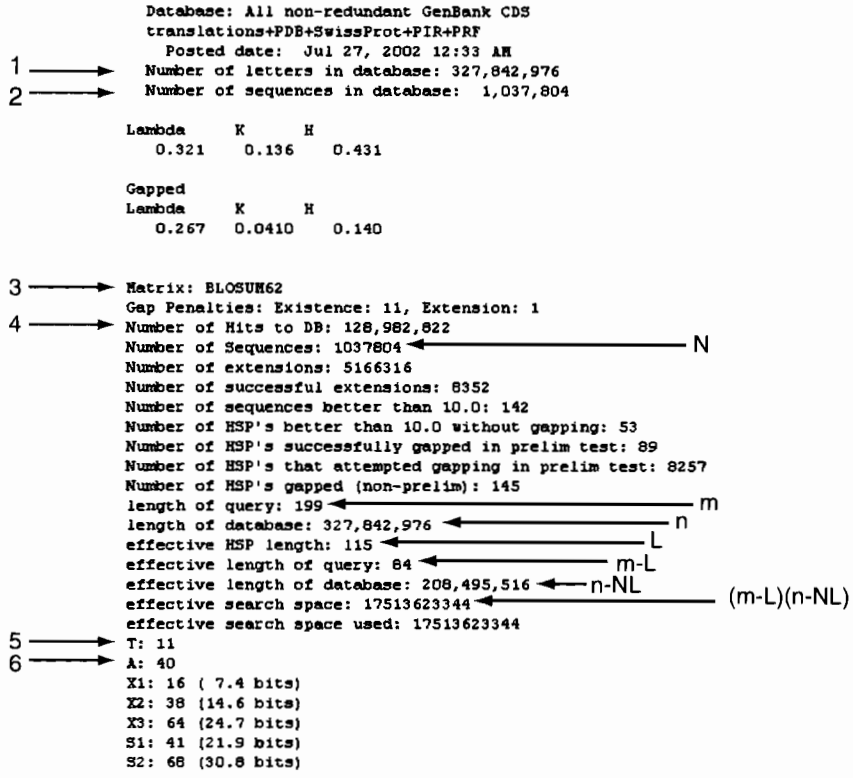


FIGURE 4.16. *BLAST search statistics. The bottom portion of the output of a blastp search using RBP4 as a query shows the name of the database, the number of residues ("letters") in the database (arrow 1), and the number of sequences (arrow 2 and letter N). Next, the λ, K, and H (entropy) values are indicated for both ungapped alignments and as estimated for gapped alignments. Beneath the matrix (arrow 3) and the gap penalty settings, the number of hits to the database are recorded (arrow 4) as well as the number of extensions, successful extensions, and HSPs better than the expect value. The length of the query (m) and the database (n) are corrected (see Equation 4.4). The effective search space is obtained by multiplying the effective length of the query by the effective length of the database. The threshold parameter (arrow 5) is indicated as well as the length that separates two independent hits to trigger an extension (arrow 6). This figure was adapted from Pagni and Jongeneel (2001). Used with permission.*

Figure 4.16 shows the parameters of a blastp search. (Similar values were presented in Table 4.3.) In the problems at the end of this chapter, we will try varying several parameters in a BLAST search. The summary statistics provide a way to evaluate how the parameters modified the search process.

BLAST Search Strategies

General Concepts

BLAST searching is a tool to explore databases of protein and DNA sequence. We have introduced the basic procedure: It is essential that you define the question you want to answer, the DNA or protein sequence you want to input, the database you want to search, and the algorithm you want to use. We will now address some basic principles regarding strategies for BLAST searching (Altschul et al., 1994). We will illustrate these issues with lipocalin and HIV-1 *pol* searches. Key issues include how to evaluate the statistical significance of BLAST search results and how to modify the optional parameters of the BLAST programs when your search yields too little or too much information. An overview of the kinds of searches that can be performed with RBP4 DNA (NM_006744) or protein (NP_006735) sequence is presented in Figure 4.17.

Principles of BLAST Searching

How to Evaluate Significance of Your Results

When you perform a BLAST search, which database matches are authentic? To answer this question, we first define a true positive as a database match that is homologous to the query sequence (descended from a common ancestor). Homology is

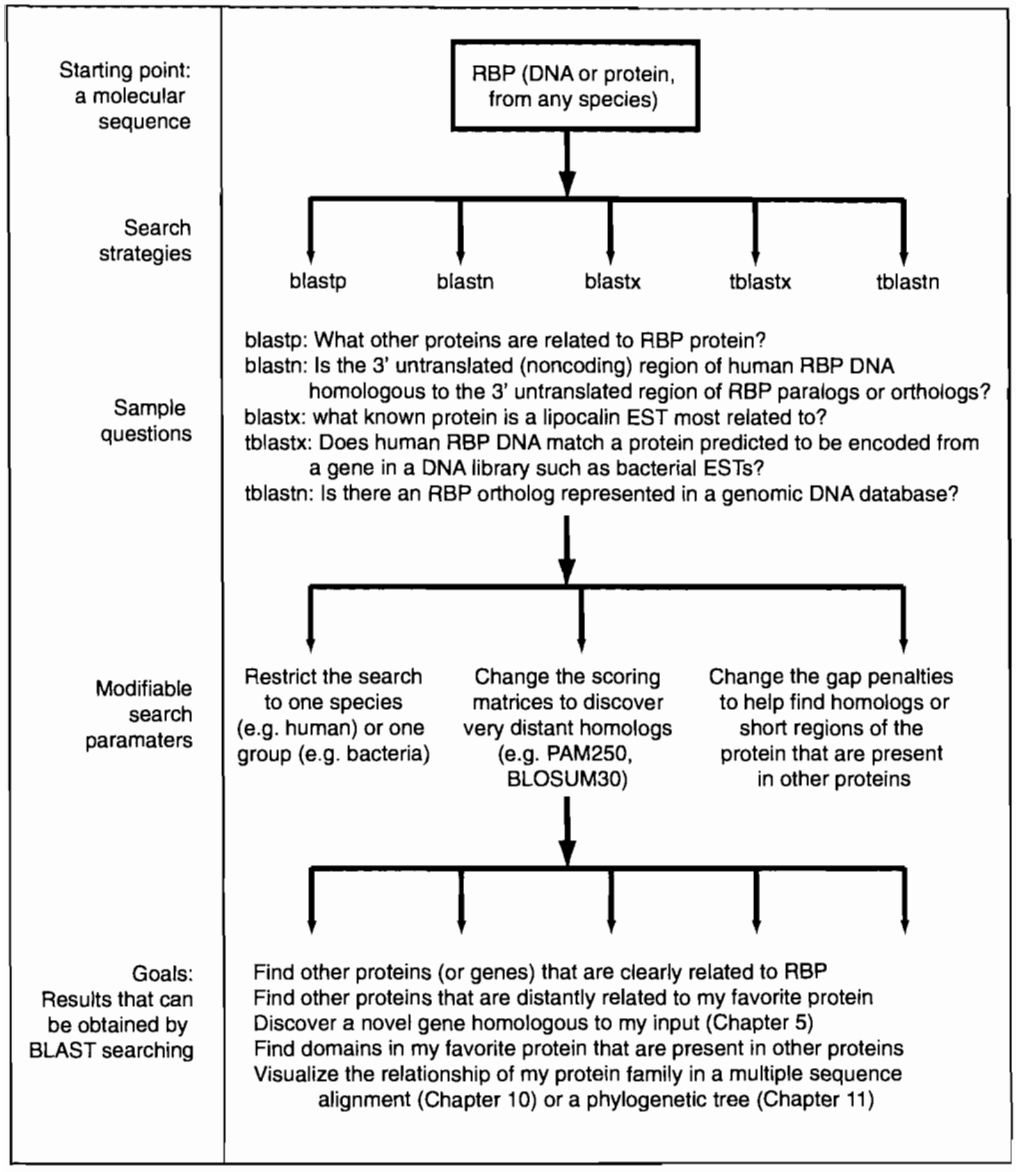


FIGURE 4.17. *Overview of BLAST searching strategies. There are many hundreds of questions that can be addressed with BLAST searching, from characterizing the genome of an organism to evaluating the sequence variation in a single gene.*

inferred based on sequence similarity, with support from statistical evaluation of the search results. The problem of assigning homology between genes or proteins is not solved by sequence analysis alone: It is also necessary to apply biological criteria to support the inference of homology. One can supplement BLAST results with evaluations of protein structure and function. The sequences of genuinely related proteins can diverge greatly, even while these proteins retain a related three-dimensional structure. Thus we expect that database searches (and pairwise protein alignments) will result in a number of false-negative matches. Many members of the lipocalin family, such as RBP4 and odorant-binding protein (OBP), share very limited sequence identity, although their three-dimensional structures are closely related and their functions of carriers of hydrophobic ligands are thought to be the same.

A consistent finding of several research groups is that the error rate of database search algorithms is reduced by using statistical scores such as expect values rather than relying on percentage identity (or percent similarity) (Brenner, 1998; Park et al., 1998; Gotoh, 1996).

Consider a blastp search of the nr database restricted to human entries using human RBP4 as a query. There are 21 entries in this case. The first way to evaluate

Sequences producing significant alignments:	Score (bits)	E Value	
gi|5803139|ref|NP_006735.1| retinol-binding protein 4, plas...	378	e-105	L
gi|20141667|sp|P02753|RETB_HUMAN Plasma retinol-binding pro...	375	e-104	L
gi|230284|pdb|1RBP| Retinol Binding Protein >gi|493897|pdb...	371	e-103	
gi|4558179|pdb|1QAB|E Chain E, The Structure Of Human Retin...	363	e-100	
gi|7770173|gb|AAF69622.1|AF119917_30 PRO2222 [Homo sapiens]	324	6e-89	L
gi|296672|emb|CAA26553.1| RBP [Homo sapiens]	207	9e-54	L
gi|5419892|emb|CAB46489.1| RBP (aa 101-172) [Homo sapiens]	149	3e-36	L
gi|2895204|gb|AAC02945.1| mutant retinol binding protein [H...	90	2e-18	L
gi|2895206|gb|AAC02946.1| mutant retinol binding protein [H...	73	3e-13	L
gi|4502163|ref|NP_001638.1| apolipoprotein D precursor [Hom...	55	5e-08	L
gi|619383|gb|AAB32200.1| apolipoprotein D, apoD [human, pla...	55	5e-08	
gi|1246096|gb|AAB35919.1| apolipoprotein D; apoD [Homo sapi...	43	3e-04	L
gi|223373|prf||0801163A complex-forming glycoprotein HC	37	0.012	
gi|21730336|pdb|1IW2|A Chain A, X-Ray Structure Of Human Co...	36	0.041	
gi|4884164|emb|CAB43305.1| hypothetical protein [Homo sapiens]	35	0.048	L
gi|130701|sp|P09466|PAEP_HUMAN Glycodelin precursor (GD) (P...	35	0.048	L
gi|4502067|ref|NP_001624.1| alpha-1-microglobulin/bikunin p...	35	0.077	L
gi|4557393|ref|NP_000597.1| complement component 8, gamma p...	34	0.16	L
gi|4505583|ref|NP_002562.1| progestagen-associated endometr...	32	0.54	L
gi|17468008|ref|XP_070794.1| similar to Glycodelin precurso...	29	3.1	L

FIGURE 4.18. *Results of a blastp nr search using human RBP as a query, restricting the output to human proteins. Note that there are 21 hits, and inspection of the E values suggests that in addition to RBP itself, several authentic paralogs may have been identified by this search. Are complement component 8 and a progestagen-associated protein likely to be homologous to RBP?*

the results is to inspect the E value list (Fig. 4.18). The nine best E value scores (from 10^{-105} to 10^{-13}) are all named RBP. This redundancy occurs because the alignments involve closely related versions of RBP. In some cases, the alignment may involve identical regions of the RBP database sequences that differ elsewhere in the sequence outside the aligned region. (For example, a fragment of a protein sequence may be deposited in the database.) Inspecting the alignment of the ninth entry (E value 3×10^{-13}) shows that it is a very short fragment of 36 amino acids that is aligned to the full-length RBP (Fig. 4.19). The next database match is apolipoprotein D (NP_001638). The RBP fragment and apolipoprotein D have very similar expect values, but in this case they have very different percent identities to RBP (94% versus 31%; Fig. 4.19). In analyzing any BLAST search results, it is important to carefully inspect the alignments as well as the scores.

Further down the list (Fig. 4.18) we see a database match of progestagen-associated endometrial protein (also called placental protein 14 and pregnancy-associated endometrial α-2-globulin; NP_002562). The alignment has a high, non-significant E value of 0.54 and a low bit score, and the protein shares only 24% amino acid identity with RBP over a span of 107 amino acid residues–including four gaps in the alignment (Fig. 4.20). One might conclude that these two proteins are not homologous. But in this case they are. In deciding whether two proteins (or DNA sequences) are homologous, one can ask several questions:

- Is the expect value significant? In this particular case it is not.
- Are the two proteins approximately the same size? It is not at all required that homologous proteins have similar sizes, and it is possible for two proteins to share only a limited domain in common. However, it is also important to develop a biological intuition about the likelihood that two proteins are homologous. A 1000-amino-acid protein with transmembrane domains is not likely to be homologous to RBP, and the vast majority of lipocalins are approximately 200 amino acids in length (20–25 kilodaltons).

```
>gi|2895206|gb|AAC02946.1|  mutant retinol binding protein [Homo sapiens]
          Length = 36

 Score = 72.8 bits (177), Expect = 3e-13
 Identities = 34/36 (94%), Positives = 35/36 (97%)

Query: 82  NWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGN 117
           NWDVCADMV TFTDTEDPAKFKMKYWGVASFLQKG+
Sbjct: 1   NWDVCADMVDTFTDTEDPAKFKMKYWGVASFLQKGS 36

>gi|4502163|ref|NP_001638.1|  apolipoprotein D precursor [Homo sapiens]
 gi|114034|sp|P05090|APOD_HUMAN   Apolipoprotein D precursor (Apo-D) (ApoD)
 gi|72088|pir||LPHUD   apolipoprotein D precursor [validated] - human
 gi|178841|gb|AAB59517.1|  apolipoprotein D precursor
 gi|178847|gb|AAA51764.1|  apolipoprotein D precursor
 gi|13938509|gb|AAH07402.1|AAH07402  apolipoprotein D [Homo sapiens]
          Length = 189

 Score = 55.5 bits (132), Expect = 5e-08
 Identities = 47/151 (31%), Positives = 78/151 (51%), Gaps = 39/151 (25%)

Query: 27  VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
           V+ENFD  ++ G WY + +K P         I A +S+ E          G++++LN  ++
Sbjct: 33  VQENFDVNKYLGRWYEI-EKIPTTFENGRCIQANYSLMEN--------GKIKVLNQ-ELR 82

Query: 87  ADMVGTFTDTE---------DPAKFKMKY-WGVASFLQKGNDDHWIVDTDYDTYAVQYSC 136
           AD  GT    E          +PAK ++K+ W + S         +WI+ TDY+ YA+ YSC
Sbjct: 83  AD--GTVNQIEGEATPVNLTEPAKLEVKFSWFMPS------APYWILATDYENYALVYSC 134

Query: 137 ----RLLNLDGTCADSYSFVFSRDPNGLPPE 163
               +L ++D      ++++ +R+PN LPPE
Sbjct: 135 TCIIQLFHVD------FAWILARNPN-LPPE 158
```

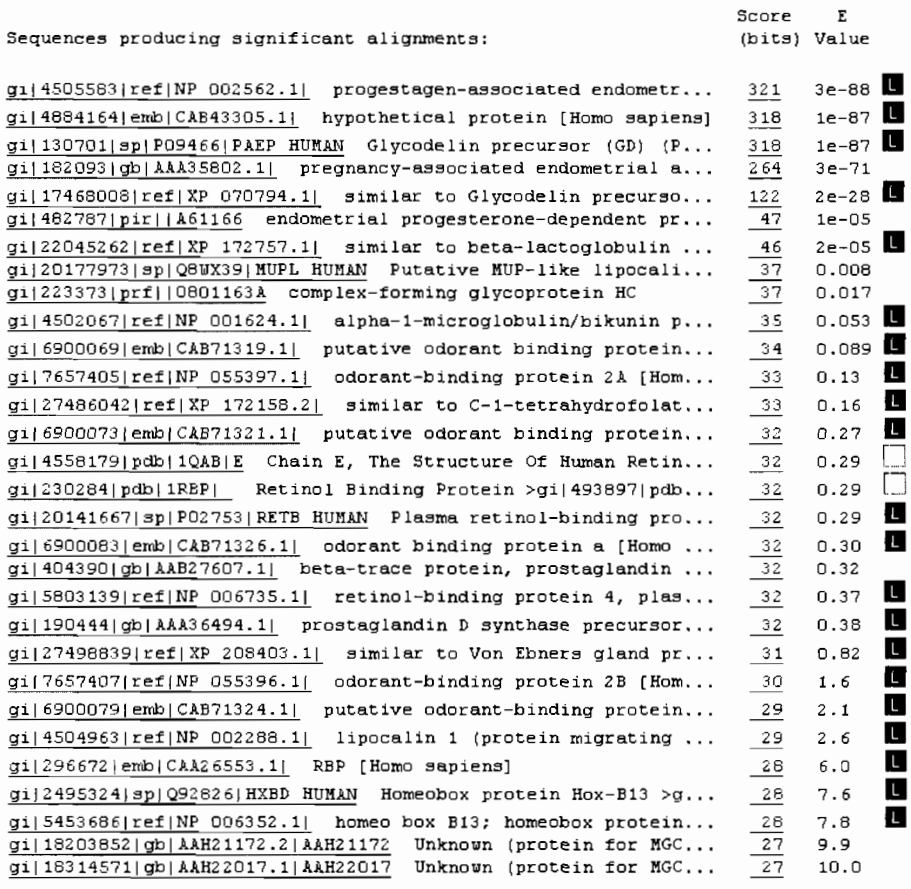
FIGURE 4.19. *Two pairwise alignments returned from the human RBP4 search (see Fig. 4.18, halfway down the list). A RBP4 fragment of just 36 amino acids yields a similar score and expect value as the longer match between RBP4 and apolipoprotein D. This result highlights the need to inspect each pairwise alignment from a BLAST search.*

- Do the proteins share a common motif or signature? In this case, both RBP4 and progestagen-associated endometrial protein have the GXW lipocalin signature. They are both part of a lipocalin multiple sequence alignment (and thus they are part of the same phylogenetic tree), as we have seen briefly already (Fig. 3.3 and 3.9).
- Are the proteins part of a reasonable multiple sequence alignment? We will see in Chapters 5 and 10 that this is the case.
- Do the proteins share a similar biological function? Like all lipocalins, both proteins are small, hydrophilic, abundant, secreted molecules.

We will define motifs and signatures in Chapter 8 and trees in Chapter 11.

```
>gi|4505583|ref|NP_002562.1|  progestagen-associated endometrial protein (placental protein 14,
          pregnancy-associated endometrial alpha-2-globulin, alpha
          uterine protein); Progestagen-associated endometrial
          protein (placental protein 14) [Homo sapiens]
 gi|190215|gb|AAA60147.1|  placental protein 14
          Length = 162

 Score = 32.0 bits (71), Expect = 0.54
 Identities = 26/107 (24%), Positives = 48/107 (44%), Gaps = 11/107 (10%)

Query: 26  RVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWD- 84
           + K++ +  + +GTW++MA      + L   + A   V  T  +          +L+ W+
Sbjct: 5   QTKQDLELPKLAGTWHSMAMAT-NNISLMATLKAPLRVHITSLLPTPEDNLEIVLHRWEN 63

Query: 85  -VCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTY 130
             C +       T +P KFK+ Y  VA       ++  ++DTDYD +
Sbjct: 64  NSCVEKKVLGEKTGNPKKFKINY-TVA-------NEATLLDTDYDNF 102
```

FIGURE 4.20. *Alignment of human RBP4 (query) with progestagen-associated endometrial protein. The bit score is relatively low, the expect value (0.49) is not significant, and in the local alignment the two proteins share only 24% amino acid identity over 107 amino acids. Nonetheless, these proteins are homologous. Their homology can be confirmed because (1) the two proteins are approximately the same size; (2) they share a lipocalin signature including a GXW motif; (3) they are multiply aligned (see Chapters 5 and 10); (4) they are both soluble, hydrophilic, abundant proteins that probably share similar functions as carrier proteins; and (6) they are very likely to share a similar three-dimensional structure (see Chapter 9).*



Sequences producing significant alignments:	Score (bits)	E Value	
gi|4505583|ref|NP_002562.1| progestagen-associated endometr...	321	3e-88	L
gi|4884164|emb|CAB43305.1| hypothetical protein [Homo sapiens]	318	1e-87	L
gi|130701|sp|P09466|PAEP_HUMAN Glycodelin precursor (GD) (P...	318	1e-87	L
gi|182093|gb|AAA35802.1| pregnancy-associated endometrial a...	264	3e-71	
gi|17468008|ref|XP_070794.1| similar to Glycodelin precurso...	122	2e-28	L
gi|482787|pir||A61166 endometrial progesterone-dependent pr...	47	1e-05	
gi|22045262|ref|XP_172757.1| similar to beta-lactoglobulin ...	46	2e-05	L
gi|20177973|sp|Q8WX39|MUPL_HUMAN Putative MUP-like lipocali...	37	0.008	
gi|223373|prf||0801163A complex-forming glycoprotein HC	37	0.017	
gi|4502067|ref|NP_001624.1| alpha-1-microglobulin/bikunin p...	35	0.053	L
gi|6900069|emb|CAB71319.1| putative odorant binding protein...	34	0.089	L
gi|7657405|ref|NP_055397.1| odorant-binding protein 2A [Hom...	33	0.13	L
gi|27486042|ref|XP_172158.2| similar to C-1-tetrahydrofolat...	33	0.16	L
gi|6900073|emb|CAB71321.1| putative odorant binding protein...	32	0.27	L
gi|4558179|pdb|1QAB|E Chain E, The Structure Of Human Retin...	32	0.29	
gi|230284|pdb|1RBP| Retinol Binding Protein >gi|493897|pdb...	32	0.29	
gi|20141667|sp|P02753|RETB_HUMAN Plasma retinol-binding pro...	32	0.29	L
gi|6900083|emb|CAB71326.1| odorant binding protein a [Homo ...	32	0.30	L
gi|404390|gb|AAB27607.1| beta-trace protein, prostaglandin ...	32	0.32	
gi|5803139|ref|NP_006735.1| retinol-binding protein 4, plas...	32	0.37	L
gi|190444|gb|AAA36494.1| prostaglandin D synthase precursor...	32	0.38	L
gi|27498839|ref|XP_208403.1| similar to Von Ebners gland pr...	31	0.82	L
gi|7657407|ref|NP_055396.1| odorant-binding protein 2B [Hom...	30	1.6	L
gi|6900079|emb|CAB71324.1| putative odorant-binding protein...	29	2.1	L
gi|4504963|ref|NP_002288.1| lipocalin 1 (protein migrating ...	29	2.6	L
gi|296672|emb|CAA26553.1| RBP [Homo sapiens]	28	6.0	L
gi|2495324|sp|Q92826|HXBD_HUMAN Homeobox protein Hox-B13 >g...	28	7.6	L
gi|5453686|ref|NP_006352.1| homeo box B13; homeobox protein...	28	7.8	L
gi|18203852|gb|AAH21172.2|AAH21172 Unknown (protein for MGC...	27	9.9	
gi|18314571|gb|AAH22017.1|AAH22017 Unknown (protein for MGC...	27	10.0	

FIGURE 4.21. *Results of a blastp nr search (restricted to human proteins) using progestagen-associated endometrial protein as a query (NP_002562). This protein is related to several proteins (complex-forming glycoprotein HC, α-1-microglobulin/bikunin) that also appear in the output of an RBP4 blastp search. This overlap supports the hypothesis that progestagen-associated endometrial protein and RBP4 are indeed homologous. It is often important to perform reciprocal BLAST searches separately using two possibly related sequences as queries.*

- Do the proteins share a similar three-dimensional structure? Although there is great diversity in lipocalin sequences, they share a remarkably well-conserved structure. This structure, a cuplike calyx, allows them to transport hydrophobic ligands across an aqueous compartment (see Chapter 9).
- If a BLAST search results in a marginal match to another protein, perform a new BLAST search using that distantly related protein as a query. A blastp nr search using progestagen-associated endometrial protein as a query results in the identification of several proteins (complex-forming glycoprotein HC and α-1-microglobulin/bikunin) that are also detected by RBP4 (Fig. 4.21). This finding increases our confidence that RBP4 and progestagen-associated endometrial protein are in fact homologous members of a protein superfamily. If the blastp search using progestagen-associated endometrial protein had shown that protein to be part of another characterized family, this would have greatly lessened our confidence that it is authentically related to RBP4.

Go to LocusLink and enter "rhodopsin," restricting the organism to human. There are over 700 loci, mostly consisting of members of this family of receptors thought to have seven transmembrane spans. We will see how to explore protein families in Chapter 8.

Historically, early database searches yielded results that were entirely unexpected. In 1984, the β-adrenergic receptor was found to be homologous to rhodopsin (Dixon et al., 1986). This was surprising because of the apparent differences between these receptors in terms of function and localization: Rhodopsin is a retina-specific receptor for light, and the adrenergic receptors were known to bind epinephrine (adrenalin) and norepinephrine, stimulating a signal transduction cascade that results in cyclic adenosine monophosphate (AMP) production. Alignment of

the protein sequences revealed that they share similar structural features (seven predicted transmembrane domains). It is now appreciated that rhodopsin and the β-adrenergic receptor are prototypic members of a superfamily of receptors that bind ligands, initiating a second messenger cascade. Another surprising finding was that some viral genes that are involved in transforming mammalian cells are actually derived from the host species. The human epidermal growth factor receptor was sequenced and found to be homologous to an avian retroviral oncogene, *v-erb-B* (Downward et al., 1984). There are many more examples of database searches that revealed unexpected relationships. We will look at an example of BLAST searching with an HIV sequence below.

How to Handle too Many Results

A common situation that is encountered in BLAST searching is that too many results are returned. There are many strategies available to limit the number of results, but to make the appropriate choices, you must focus on the question you are trying to answer.

- Use the "limit by Entrez query" window to enter "refseq" and all the hits that are returned will have RefSeq accession numbers. This will often eliminate redundant database matches.
- Limit the database returns by organism, when applicable. This may eliminate extraneous information. If you use the options feature of the BLAST server to limit a search by organism, the same size search is performed; in contrast, if you choose an organism-specific database, this often increases the speed of the search dramatically. We present some organism-specific BLAST servers in Chapter 5.
- Use just a portion of the query sequence, when appropriate. A search of a multidomain protein can be performed with just the isolated domain sequence. If you are studying HIV-1 pol, you may be interested in the entire protein or in a specific portion such as the reverse transcriptase domain.
- Adjust the scoring matrix to make it more appropriate to the degree of similarity between your query and the database matches.
- Adjust the expect value; lowering E reduces the number of database matches that are returned.

How to Handle too Few Results

Many genes and proteins have no significant database matches or have very few. As new microbial genomes are sequenced, as many as half the predicted proteins have no matches to any other proteins (Chapter 14). Some strategies to increase the number of database matches from BLAST searching are obvious: remove Entrez limits, raise the expect values, and try scoring matrices with higher PAM or lower BLOSUM values. One can also search a large variety of additional databases. Within the NCBI website, one can search all available databases (e.g., HTGS and GSS). Many genome-sequencing centers for a variety of organisms maintain separate databases that can be searched by BLAST. These are described in Chapter 5 (advanced BLAST searching). Additionally, there are many database searching algorithms that are more sensitive than BLAST. These include position-specific scoring matrices (PSSMs) and hidden Markov models (HMMs) and are also described in Chapters 5 and 10.

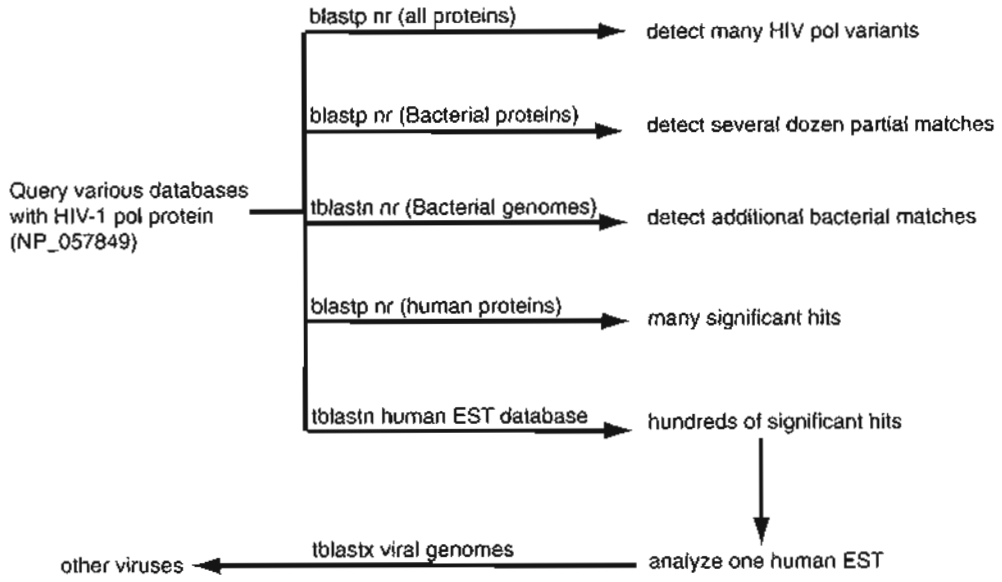


FIGURE 4.22. Overview of BLAST searches beginning with HIV-1 pol protein. A series of BLAST searches can often be performed to pursue questions about a particular gene, protein, or organism. The number of database matches returned by a BLAST search can vary from none to thousands and depends entirely on the nature of the query, the database, and the search parameters.

BLAST Searching with Multidomain Protein: HIV-1 pol

The pol protein of HIV-1 (NP_057849) has 1003 amino acid residues and includes separate protease, reverse transcriptase, and integrase domains. Thus it is an example of a multidomain protein. Figure 4.22 previews the kinds of searches we can perform with a viral protein such as this one.

What happens upon BLAST searching the nonredundant protein database with this protein? We input the RefSeq accession number (NP_057849) and click submit. The program reports that putative conserved domains have been detected, and a schematic of the protein indicates the location of each domain (Fig. 4.23). Clicking on any of these domains links to the NCBI conserved domain database as well as to the Pfam and SMART databases (see Chapters 5 and 8). Continuing with the BLAST search, we see that there are 500 hits, all with extremely low expect values, and all correspond to HIV and simian immunodeficiency virus (SIV) pol protein matches

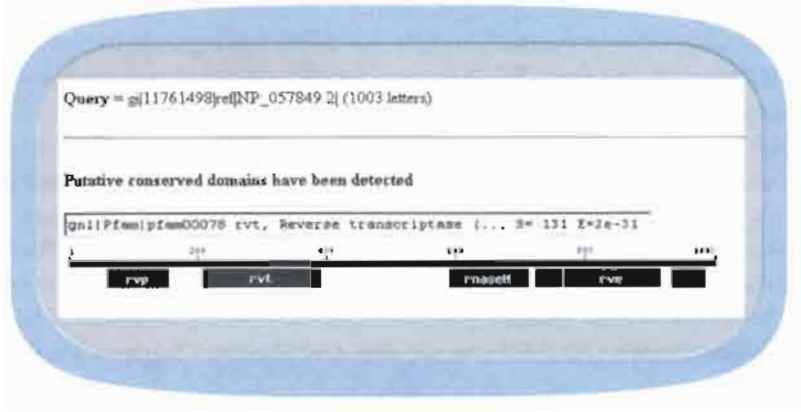


FIGURE 4.23. A blastp search with viral pol *(NP_057849) shows conserved domains in the protein. These blocks are clickable and link to additional databases (Pfam, SMART; see Chapter 8). The abbreviations are* rvp, *retroviral aspartyl protease;* rvt, *reverse transcriptase (RNA-dependent DNA polymerase);* rnaseH, *ribonuclease H; small unlabeled box, integrase zinc domain;* rve, *integrase core domain; small unlabeled box, integrase DNA-binding domain.*

Sequences producing significant alignments:	Score (bits)	E Value
gi|28872819|ref|NP_057849.4| Gag-Pol; Gag-Pol polyprotein [...	2836	0.0
gi|11066865|gb|AAG28737.1|AF287353_1 gag-pol fusion protein...	2756	0.0
gi|3002831|gb|AAD03191.1| gag-pol fusion polyprotein [Human...	2753	0.0
gi|19072104|dbj|BAB85751.1| Gag-pol fusion polyprotein [Hum...	2744	0.0
gi|3002841|gb|AAD03200.1| gag-pol fusion polyprotein [Human...	2731	0.0
gi|11095912|gb|AAG30116.1|AF286365_2 gag-pol fusion polypro...	2723	0.0
gi|3002860|gb|AAD03217.1| gag-pol fusion polyprotein [Human...	2716	0.0
gi|3193273|gb|AAD03326.1| gag-pol fusion polyprotein [Human...	2691	0.0
gi|3002887|gb|AAD03241.1| gag-pol fusion polyprotein [Human...	2688	0.0
gi|3002869|gb|AAD03225.1| gag-pol fusion polyprotein [Human...	2685	0.0
gi|3002851|gb|AAD03209.1| gag-pol fusion polyprotein [Human...	2674	0.0
gi|3002878|gb|AAD03233.1| gag-pol fusion polyprotein [Human...	2667	0.0
gi|25166689|gb|AAN73492.1|AF484483_1 gag-pol fusion polypro...	2647	0.0
gi|25166961|gb|AAN73736.1|AF484511_1 gag-pol fusion polypro...	2640	0.0
gi|25166629|gb|AAN73438.1|AF484477_1 gag-pol fusion polypro...	2633	0.0
gi|25166991|gb|AAN73763.1|AF484514_1 gag-pol fusion polypro...	2632	0.0
gi|25166901|gb|AAN73682.1|AF484505_1 gag-pol fusion polypro...	2630	0.0
gi|25166951|gb|AAN73727.1|AF484510_1 gag-pol fusion polypro...	2628	0.0
gi|25166731|gb|AAN73529.1|AF484488_1 gag-pol fusion polypro...	2627	0.0
gi|25167011|gb|AAN73781.1|AF484516_1 gag-pol fusion polypro...	2627	0.0
gi|25166791|gb|AAN73583.1|AF484494_1 gag-pol fusion polypro...	2627	0.0
gi|25166821|gb|AAN73610.1|AF484497_1 gag-pol fusion polypro...	2626	0.0
gi|25166711|gb|AAN73511.1|AF484486_1 gag-pol fusion polypro...	2625	0.0
gi|25166751|gb|AAN73547.1|AF484490_1 gag-pol fusion polypro...	2624	0.0
gi|25166669|gb|AAN73474.1|AF484481_1 gag-pol fusion polypro...	2622	0.0
gi|25166981|gb|AAN73754.1|AF484513_1 gag-pol fusion polypro...	2622	0.0
gi|25166801|gb|AAN73592.1|AF484495_1 gag-pol fusion polypro...	2622	0.0
gi|25166891|gb|AAN73673.1|AF484504_1 gag-pol fusion polypro...	2622	0.0
gi|25167041|gb|AAN73808.1|AF484519_1 gag-pol fusion polypro...	2621	0.0
gi|25166911|gb|AAN73691.1|AF484506_1 gag-pol fusion polypro...	2620	0.0
gi|25166841|gb|AAN73628.1|AF484499_1 gag-pol fusion polypro...	2620	0.0
gi|25166741|gb|AAN73538.1|AF484489_1 gag-pol fusion polypro...	2619	0.0
gi|25166831|gb|AAN73619.1|AF484498_1 gag-pol fusion polypro...	2619	0.0
gi|18073999|emb|CAC86564.1| gag-pol fusion polyprotein prec...	2618	0.0
gi|25166871|gb|AAN73655.1|AF484502_1 gag-pol fusion polypro...	2617	0.0
gi|2570314|gb|AAC97575.1| gag-Pol fusion [Human immunodefic...	2615	0.0
gi|25166679|gb|AAN73483.1|AF484482_1 gag-pol fusion polypro...	2614	0.0
gi|25166771|gb|AAN73565.1|AF484492_1 gag-pol fusion polypro...	2613	0.0
gi|25166701|gb|AAN73502.1|AF484485_1 gag-pol fusion polypro...	2613	0.0
gi|25166851|gb|AAN73637.1|AF484500_1 gag-pol fusion polypro...	2612	0.0
gi|25167071|gb|AAN73835.1|AF484522_1 gag-pol fusion polypro...	2608	0.0
gi|4324862|gb|AAD17126.1| gag-pol polyprotein [Human immuno...	2604	0.0
gi|25167021|gb|AAN73790.1|AF484517_1 gag-pol fusion polypro...	2603	0.0
gi|25166721|gb|AAN73520.1|AF484487_1 gag-pol fusion polypro...	2603	0.0
gi|25166659|gb|AAN73465.1|AF484480_1 gag-pol fusion polypro...	2600	0.0
gi|25167001|gb|AAN73772.1|AF484515_1 gag-pol fusion polypro...	2597	0.0

FIGURE 4.24. *A blastp nr search with HIV-1* pol *results in a very large number of database hits that all appear to be variants of HIV-1. Note that all the E values shown are zero. This result obscures any possible hits that are not from HIV-1.*

(Fig. 4.24). Changing the output format to "query-anchored without identities" is one way to view the dramatic conservation of these viral proteins (Fig. 4.25).

To learn more about the distribution of pol proteins throughout the tree of life, we may further ask what bacterial proteins are related to the viral HIV-1 pol polyprotein. Repeat the blastp search with NP_057849 as the query, but select "Bacteria" from the pull-down menu of organisms. Here, the graphical overview of BLAST search results is extremely helpful to show that two domains of viral pol have the majority of matches to known bacterial sequences, corresponding to amino acids 600–700 and 800–900 of pol (Fig. 4.26). Comparison of this output to the domain architecture of HIV-1 pol (Fig. 4.23) suggests that the two viral protein domains with matches to bacterial proteins are RNAse H and an integrase. Indeed, the bacterial matches aligned to viral pol include ribonuclease H and transposases (Fig. 4.27a). Inspection of the pairwise alignments indicates that the viral and bacterial proteins are homologous, sharing up to 30% amino acid identity over spans of up to 166 amino acids (Fig. 4.27*b*).

Let us now turn our attention to human proteins that may be homologous to HIV-1 pol. The search is identical to our search of bacteria, except that we restrict the organism to *Homo sapiens*. Interestingly, there are over 50 human matches, and

```
1_19000    294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 353
28872819   294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 353
11066865   294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 353
3002831    294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGV 353
19072104   294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 353
3002841    295  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 354
11095912   294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGV 353
3002860    294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNSNPDCKTILKALGPAATLEEMMTACQGV 353
3193273    294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNSNPDCKTILKALGPGATLEEMMTACQGV 353
3002887    299  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 358
3002869    294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 353
3002851    299  RDYVDRFYKTLRAEQATQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 358
3002878    294  RDYVDRFYKTLRAEQASQEVKNWMTETLLVQNANPDCRTILKALGPAATLKEMMTACQGV 353
25166689   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGV 351
25166961   292  RDYVDRFYKTLRAEQATQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166629   292  RDYVDRFYKTLRAEQASQEVKGWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166991   292  RDYVDRFYKTLRAEQASQEVKSWMTETLLVQNANPDCKTILKSLGTGATLEEMMTACQGV 351
25166901   292  RDYVDRFYKTLRAEQATQEVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166951   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166731   291  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 350
25167011   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166791   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLIQNANPDCKTILKALGPGATLEEMMTACQGV 351
25166821   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCRTILKALGPAATLEEMMTACQGV 351
25166711   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166751   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166669   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166981   293  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPGASLEEMMTACQGV 352
25166801   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166891   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGV 351
25167041   292  RDYVDRFYKTLRAEQATQEVKNWMTETLLIQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166911   292  RDYVDRFYKTLRAEQASQDVKTWMTETLLVQNANPDCKTILKALGPQATLEEMMTACQGV 351
25166841   290  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 349
25166741   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166831   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
18073999   294  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGV 353
25166871   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCRTILKALGPGATLEEMMTACQGV 351
2570314    294  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 353
25166679   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166771   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166701   292  RDYVDRFYKTLRAEQASQEVKGWMTETLLVQNANPDCKTILKALGPGATLEEMMTACQGV 351
25166851   292  RDYVDRFFKILRAEQATQEVKNWMTDTLLVQNANPDCKTILKALGPAASLEEMMTACQGV 351
25167071   292  RDYVDRFFKTLRAEQATQEVKGWMTDTLLVQNANPDCKTILKALGPGASLEEMMTACQGV 351
4324862    291  RDYVDRFFKTLRAEQATQDVKNWMTDTLLVQNANPDCKTILRALGPGATLEEMMTACQGV 350
25167021   292  RDYVDRFYKTLRAEQASQEVKNWMTETLLIQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166721   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAATLEEMMTACQGV 351
25166659   292  RDYVDRFYKTLRAEQASQDVKNWMTETLLVQNANPDCKTILKALGPAASLEEMMTACQGV 351
```

FIGURE 4.25. *Portion of the output of a blastp search using the HIV-1 pol protein as a query (NP_057849). The flat query-anchored output format reveals substituted amino acid residues as well as those that remain invariant.*

a number of these span the full length of viral pol (Fig. 4.28). These human proteins have been annotated as polymerases (Fig. 4.28*b*). Are these human genes expressed? If so, they should produce RNA transcripts that may be characterized as ESTs from cDNA libraries. Perform a search of human ESTs with the viral pol protein; it is necessary to use the translating BLAST website with the tblastn algorithm, and the database must be set to EST (Fig. 4.29). There are hundreds of human transcripts, actively transcribed, that are predicted to encode proteins homologous to viral pol (Fig. 4.30). In Chapter 6, we will see how to evaluate these human ESTs to determine where in the body they are expressed and when during development they are expressed. Some researchers have suggested that neuropsychiatric diseases such as schizophrenia are associated with elevated levels of endogenous retroviral gene expression (Karlsson et al., 2001).

Could the human ESTs that are homologous to HIV-1 *pol* be even more closely related to other viral *pol* genes? To answer this question, select a human EST that

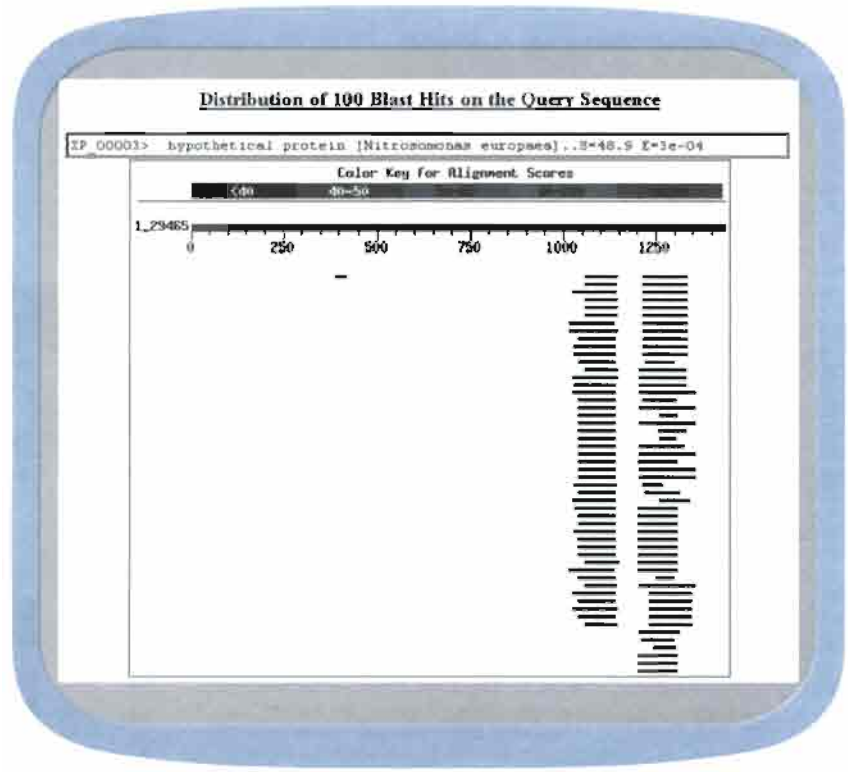


FIGURE 4.26. Result of a blastp search with HIV-1 pol *as a query, restricting the output to bacteria. The graphical output of the BLAST search allows easy identification of the domains within HIV-1 that have bacterial matches. (Refer to Fig. 4.23.)*

we found is related to HIV-1 *pol* (from Fig. 4.30; we will choose accession AI636743). Perform the computationally intensive tblastx search using this EST's accession as an input and restrict the organism of the search to viruses. At the present time, this search results in the identification of many endogenous human retroviral sequences (i.e., DNA sequences that are part of the human genome) that are mistakenly left in the viruses division of GenBank. The virus that is most closely related to this particular human EST is a virus that afflicts sheep (accession AF105220). We initially performed a BLAST search with an HIV query and have used a further series of BLAST searches to gain insight into the biology of HIV-1 *pol*.

BLAST Searching with Lipocalins: Effect of Changing Scoring Matrices

Scoring matrices can have dramatic effects on BLAST search results, just as they can affect pairwise alignments (Chapter 3). We performed a blastp nr search restricted to human proteins using "human RBP4" as a query (Fig. 4.18). The default scoring matrix was BLOSUM62. Using a PAM30 matrix that is appropriate for relatively closely related proteins, fewer results are returned, and a match such as apolipoprotein D receives an expect value of 10^{-4} rather than 10^{-8} (Fig. 4.31*a*). Using a PAM70 scoring matrix, additional lipocalin results are evident (Fig. 4.31*b*). A search of human proteins using rat OBP as a query results in only five matches (of which only three are authentic lipocalins) (Fig. 4.32). Changing the matrix to BLOSUM45, which tends to permit detection of proteins sharing less amino acid identity, there are now 12 matches, including four lipocalins.

(*a*)

Sequences producing significant alignments:		Score (bits)	E Value	
gi|22955850|gb|ZP_00003651.1|	hypothetical protein [Nitroso...	49	3e-04	
gi|22954454|gb|ZP_00002255.1|	hypothetical protein [Nitroso...	49	3e-04	
gi|22954317|gb|ZP_00002118.1|	hypothetical protein [Nitroso...	49	3e-04	
gi|22954687|gb|ZP_00002488.1|	hypothetical protein [Nitroso...	49	3e-04	
gi|22956290|gb|ZP_00004091.1|	hypothetical protein [Nitroso...	49	3e-04	
gi|22955780|gb|ZP_00003581.1|	hypothetical protein [Nitroso...	49	3e-04	
gi|22955002|gb|ZP_00002803.1|	hypothetical protein [Nitroso...	49	3e-04	
gi|22954366|gb|ZP_00002167.1|	hypothetical protein [Nitroso...	49	3e-04	
gi|22955688|gb|ZP_00003489.1|	hypothetical protein [Nitroso...	49	3e-04	
gi|22956184|gb|ZP_00003985.1|	hypothetical protein [Nitroso...	49	3e-04	
gi|22954280|gb|ZP_00002081.1|	hypothetical protein [Nitroso...	49	3e-04	
gi|22954184|gb|ZP_00001985.1|	hypothetical protein [Nitroso...	43	0.012	
gi|23103018|ref|ZP_00089511.1|	hypothetical protein [Azotob...	42	0.042	
gi|23469154|ref|ZP_00124489.1|	hypothetical protein [Pseudo...	41	0.065	
gi|17546232|ref|NP_519634.1|	PROBABLE RIBONUCLEASE HI PROTE...	41	0.067	
gi|28870867|ref|NP_793486.1|	ribonuclease HI [Pseudomonas s...	41	0.071	
gi|13474791|ref|NP_106361.1|	transposase [Mesorhizobium lot...	40	0.078	
gi|26990835|ref|NP_746260.1|	ribonuclease HI [Pseudomonas p...	40	0.11	
gi|22978683|gb|ZP_00024428.1|	hypothetical protein [Ralston...	39	0.17	
gi|22977454|gb|ZP_00023257.1|	hypothetical protein [Ralston...	39	0.17	
gi|2760874|gb|AAD10223.1|	ORFB [Xanthomonas campestris pv. ...	39	0.18	
gi|22978013|gb|ZP_00023785.1|	hypothetical protein [Ralston...	39	0.18	
gi|15805924|ref|NP_294623.1|	ribonuclease H [Deinococcus ra...	39	0.23	
gi|15597012|ref|NP_250506.1|	ribonuclease H [Pseudomonas ae...	39	0.27	
gi|24111645|ref|NP_706155.1|	RNase HI, degrades RNA of DNA-...	38	0.43	
gi|23106231|ref|ZP_00092685.1|	hypothetical protein [Azotob...	38	0.45	
gi|494541|pdb|1RBS|	Ribonuclease H (E.C.3.1.26.4) Mutant W...	38	0.47	S
gi|13787086|pdb|1G15|A	Chain A, Co-Crystal Of E. Coli Rnase...	38	0.53	S
gi|23061138|ref|ZP_00086001.1|	hypothetical protein [Pseudo...	38	0.54	
gi|12084647|pdb|1F21|A	Chain A, Divalent Metal Cofactor Bin...	38	0.57	S
gi|20150576|pdb|1JXB|A	Chain A, I53a, A Point Mutant Of The...	37	0.64	S
gi|4378703|gb|AAD19632.1|	unknown [Xanthomonas campestris p...	37	0.64	
gi|23041530|ref|ZP_00072980.1|	hypothetical protein [Tricho...	37	0.67	
gi|494049|pdb|1GOA|	Ribonuclease H (E.C.3.1.26.4) Mutant W...	37	0.79	S
gi|15799890|ref|NP_285902.1|	RNase HI, degrades RNA of DNA-...	37	0.82	S
gi|443224|pdb|1RDA|	Ribonuclease H (E.C.3.1.26.4) Mutant W...	37	0.84	S
gi|27479697|gb|AAO17224.1|	RnhA [Photorhabdus luminescens]	37	0.90	
gi|21910590|ref|NP_664858.1|	putative transposase [Streptoc...	37	0.92	
gi|20150465|pdb|1JL1|A	Chain A, D10a E. Coli Ribonuclease Hi	37	0.92	S
gi|494540|pdb|1RBR|	Ribonuclease H (E.C.3.1.26.4) Mutant W...	37	0.95	S
gi|494543|pdb|1RBU|	Ribonuclease H (E.C.3.1.26.4) Mutant W...	37	1.0	S
gi|443089|pdb|1LAV|	Ribonuclease H (E.C.3.1.26.4) Mutant W...	37	1.1	S
gi|230303|pdb|1RNH|	Selenomethionyl Ribonuclease H (E.C.3....	37	1.1	S
gi|443090|pdb|1LAW|	Ribonuclease H (E.C.3.1.26.4) Mutant W...	37	1.1	S

(*b*)

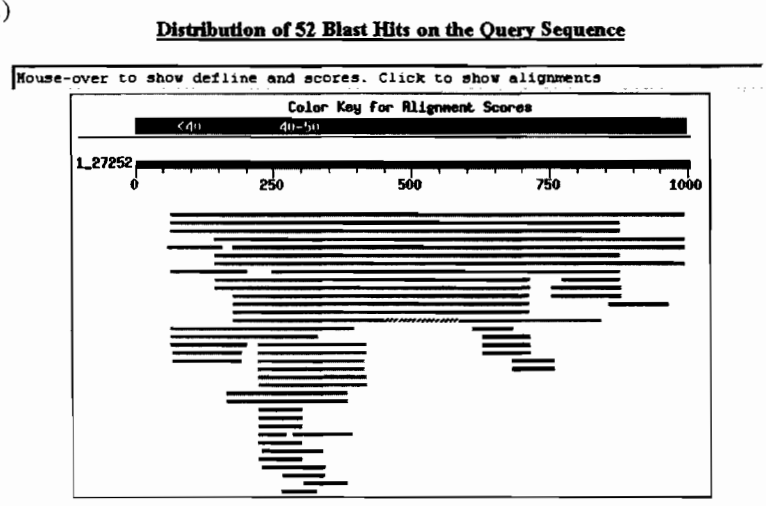
☐ >gi|22955850|gb|ZP_00003651.1| hypothetical protein [Nitrosomonas europaea]
Length = 320

Score = 48.9 bits (115), Expect = 3e-04
Identities = 44/134 (32%), Positives = 60/134 (44%), Gaps = 14/134 (10%)

```
Query: 1216 EGKVIL-VAVHVASGYIEAEVIPAETGQETAYFLLKLAGRWPVK--TIHTDNGSNFTG-A 1271
          EGK+ L VA+   S +  AE++P     E A FL  L    P K  TI TDNG  FT
Sbjct: 146  EGKLYLFVAIDRTSKFAYAELLPKYGKMEAAQFLRNLIAAVPYKIHTILTDNGIQFTHRK 205

Query: 1272 TVRAA--------CWWAGIKQEFGIPYNPQSQGVVESMNKELKKIIGQVRDQAEHLKTAV 1323
          T R A         C   G +     P +P + G VE MN+ LK+   +      H +
Sbjct: 206  TDRHAFLHIFDRVCLENGTEHRLTQPNHPWTNGQVERMNRTLKEATVKRYHYENHQQLRE 265

Query: 1324 QMAVFI--HNFKRK 1335
           +  F+  +NF R+
Sbjct: 266  HLYSFLNAYNFARR 279
```

FIGURE 4.27. *(a) Bacterial proteins that are identified in a blastp search with HIV-1 pol include transposases and ribonuclease H proteins. (b) Alignment of a bacterial protein with an HIV-1* pol *query suggests that the sequences are homologous, having signficant expect values and about 30% amino acid identity over extended regions of the alignments.*

(*a*)



(*b*)

Sequences producing significant alignments:		Score (bits)	E Value
gi|5802821|gb|AAD51797.1|AF164614_1	(AF164614) Gag-Pro-Pol ...	239	2e-62
gi|5802814|gb|AAD51793.1|AF164611_1	(AF164611) Gag-Pro-Pol-...	238	2e-62
gi|5802819|gb|AAD51796.1|AF164613_1	(AF164613) Gag-Pro-Pol ...	238	2e-62
gi|13021822|gb|AAK11553.1|AF298587_1	(AF298587) polymerase ...	235	2e-61
gi|13021825|gb|AAK11554.1|AF298588_1	(AF298588) polymerase ...	235	2e-61
gi|130393|sp|P10266|POL1_HUMAN	RETROVIRUS-RELATED POL POLYP...	232	2e-60
gi|4456990|gb|AAD21097.1|	(AF074086) polymerase [Homo sapie...	232	2e-60
gi|1196429|gb|AAA88033.1|	(M14123) pol/env ORF (bases 3878-...	190	7e-48
gi|231033|pdb|4HVP|A	Chain A, HIV-1 Protease (HIV-1 PR) Com...	189	2e-47
gi|3600064|gb|AAC63290.1|	(AF080229) polymerase [Homo sapiens]	187	7e-47
gi|3600067|gb|AAC63291.1|	(AF080231) polymerase [Homo sapiens]	182	2e-45
gi|3600073|gb|AAC63294.1|	(AF080234) polymerase [Homo sapiens]	177	4e-44
gi|3600071|gb|AAC63293.1|	(AF080233) polymerase [Homo sapiens]	177	9e-44
gi|3600069|gb|AAC63292.1|	(AF080232) polymerase [Homo sapiens]	175	3e-43
gi|10504249|gb|AAG18012.1|	(AF248269) gag-pro-pol precursor...	145	2e-34
gi|5802810|gb|AAD51791.1|	(AF164609) Gag-Pro-Pol protein [H...	139	1e-32
gi|5802824|gb|AAD51799.1|AF164615_1	(AF164615) Gag-Pro-Pol ...	125	2e-28
gi|2257764|gb|AAB63113.1|	(U87590) polymerase [Homo sapiens]	118	3e-26
gi|2257766|gb|AAB63114.1|	(U87591) polymerase [Homo sapiens]	117	7e-26
gi|2257762|gb|AAB63112.1|	(U87589) polymerase [Homo sapiens]	111	4e-24
gi|2257773|gb|AAB63117.1|	(U87595) polymerase [Homo sapiens]	111	6e-24
gi|2257770|gb|AAB63116.1|	(U87593) polymerase [Homo sapiens]	107	7e-23
gi|2257768|gb|AAB63115.1|	(U87592) polymerase [Homo sapiens]	105	3e-22
gi|14760974|ref|XP_034809.1|	similar to putative gag-pro-po...	77	2e-13

FIGURE 4.28. *(a) Graphical output of a blastp search using HIV-1 pol protein to search for matches against human proteins. (b) Note that there are many more human hits than there were bacterial matches.*

PERSPECTIVE

BLAST searching has emerged as an indispensable tool to analyze the relation of a DNA or protein sequence to millions of sequences in public databases. All database search tools confront the issues of sensitivity (i.e., the ability to minimize false-negative results), selectivity (i.e., the ability to minimize false-positive results), and time. As the size of the public databases has grown exponentially in recent years, the BLAST tools have evolved to provide a rapid, reliable way to screen the databases.

PITFALLS

There are several common pitfalls to avoid in BLAST searching. The most common error among novice BLAST users is to search protein or DNA sequences against the wrong database. It is also important to understand the basic BLAST algorithms. These concepts are summarized in Figure 4.3.

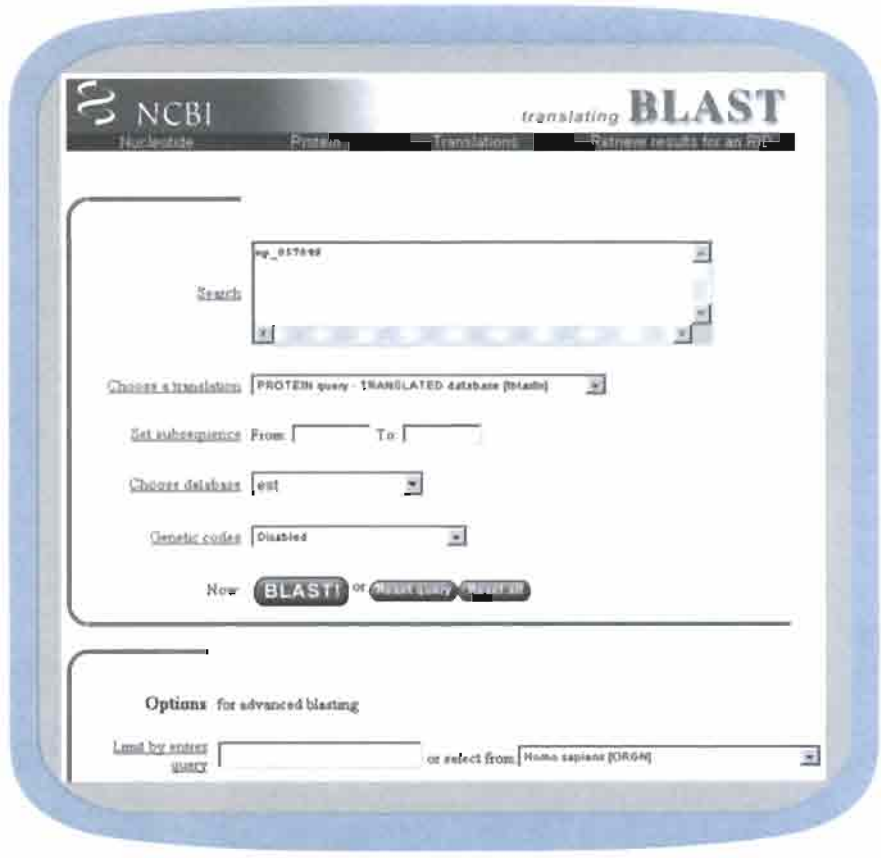


FIGURE 4.29. Are human transcripts expressed that encode proteins homologous to HIV-1 pol *protein? To answer this question, set up a tblastn search against a* H. sapiens *EST database and click "BLAST!"*

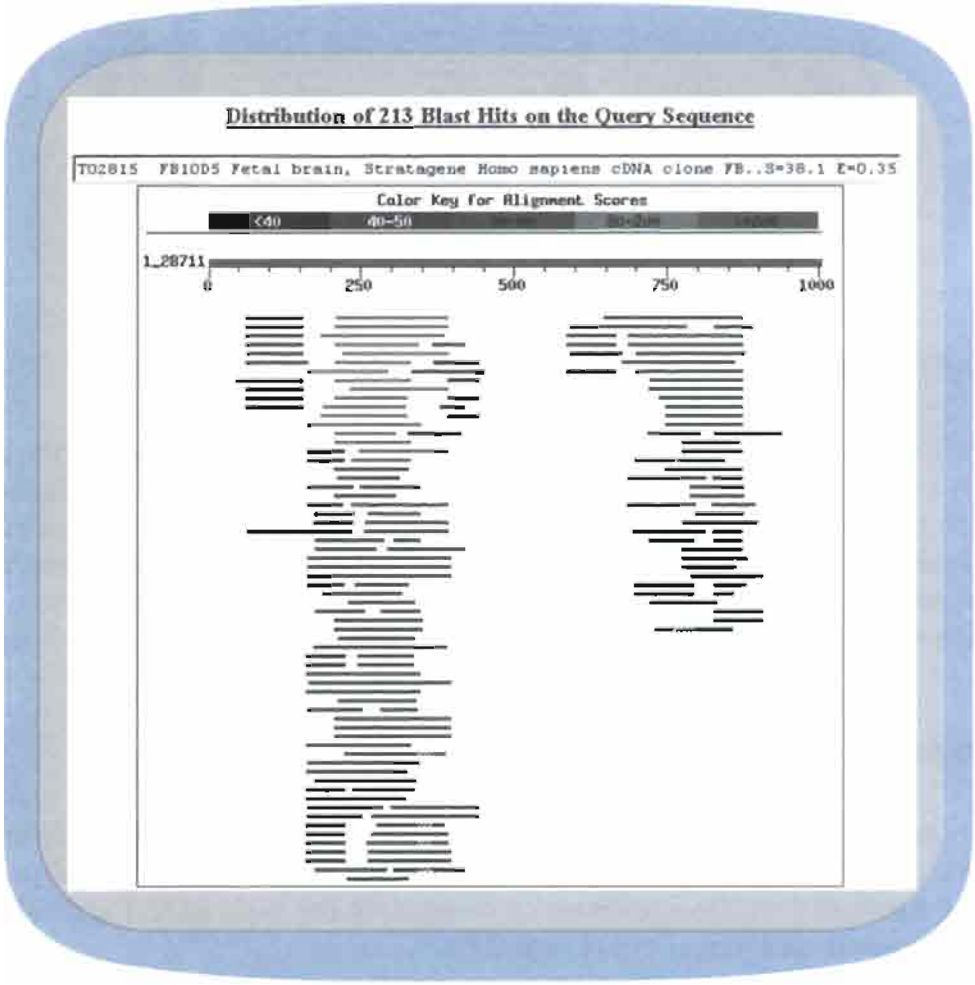


FIGURE 4.30. Results of a tblastn search with viral pol *protein against a human EST database, showing that many human genes are actively transcribed to generate transcripts predicted to make proteins homologous to HIV-1* pol.

(a)

Sequences producing significant alignments:	Score (bits)	E Value	
gi|20141667|sp|P02753|RETB_HUMAN Plasma retinol-binding pro...	578	e-165	L
gi|5803139|ref|NP_006735.1| retinol-binding protein 4, plas...	576	e-164	L
gi|230284|pdb|1RBP| Retinol Binding Protein >gi|493897|pdb...	574	e-164	S
gi|4558179|pdb|1QAB|E Chain E, The Structure Of Human Retin...	554	e-158	S
gi|7770173|gb|AAF69622.1|AF119917_30 PRO2222 [Homo sapiens]	498	e-141	L
gi|296672|emb|CAA26553.1| RBP [Homo sapiens]	313	3e-85	L
gi|5419892|emb|CAB46489.1| RBP (aa 101-172) [Homo sapiens]	226	4e-59	L
gi|2895204|gb|AAC02945.1| mutant retinol binding protein [H...	138	2e-32	L
gi|2895206|gb|AAC02946.1| mutant retinol binding protein [H...	109	6e-24	L
gi|619383|gb|AAB32200.1| apolipoprotein D, apoD [human, pla...	45	2e-04	
gi|1246096|gb|AAB35919.1| apolipoprotein D; apoD [Homo sapi...	44	3e-04	L
gi|4502163|ref|NP_001638.1| apolipoprotein D precursor [Hom...	44	3e-04	L
gi|4506251|ref|NP_000945.1| prostaglandin D2 synthase 21kDa...	32	1.0	L
gi|404390|gb|AAB27607.1| beta-trace protein, prostaglandin ...	32	1.7	
gi|190444|gb|AAA36494.1| prostaglandin D synthase precursor...	32	1.7	L
gi|27480905|ref|XP_209977.1| similar to CALII [Gallus gallu...	30	4.2	L

(b)

Sequences producing significant alignments:	Score (bits)	E Value	
gi|20141667|sp|P02753|RETB_HUMAN Plasma retinol-binding pro...	498	e-141	L
gi|5803139|ref|NP_006735.1| retinol-binding protein 4, plas...	496	e-140	L
gi|230284|pdb|1RBP| Retinol Binding Protein >gi|493897|pdb...	495	e-140	S
gi|4558179|pdb|1QAB|E Chain E, The Structure Of Human Retin...	480	e-135	S
gi|7770173|gb|AAF69622.1|AF119917_30 PRO2222 [Homo sapiens]	430	e-120	L
gi|296672|emb|CAA26553.1| RBP [Homo sapiens]	269	5e-72	L
gi|5419892|emb|CAB46489.1| RBP (aa 101-172) [Homo sapiens]	196	3e-50	L
gi|2895204|gb|AAC02945.1| mutant retinol binding protein [H...	119	6e-27	L
gi|2895206|gb|AAC02946.1| mutant retinol binding protein [H...	97	4e-20	L
gi|619383|gb|AAB32200.1| apolipoprotein D, apoD [human, pla...	61	3e-09	
gi|4502163|ref|NP_001638.1| apolipoprotein D precursor [Hom...	59	6e-09	L
gi|1246096|gb|AAB35919.1| apolipoprotein D; apoD [Homo sapi...	50	4e-06	L
gi|223373|prf||0801163A complex-forming glycoprotein HC	38	0.019	
gi|4502067|ref|NP_001624.1| alpha-1-microglobulin/bikunin p...	35	0.16	L
gi|4506251|ref|NP_000945.1| prostaglandin D2 synthase 21kDa...	31	2.1	L
gi|404390|gb|AAB27607.1| beta-trace protein, prostaglandin ...	30	3.1	
gi|190444|gb|AAA36494.1| prostaglandin D synthase precursor...	30	3.1	L
gi|27480905|ref|XP_209977.1| similar to CALII [Gallus gallu...	30	3.6	L
gi|6841194|gb|AAF28950.1|AF161390_1 HSPC272 [Homo sapiens]	29	5.7	L
gi|2224565|dbj|BAA20771.1| KIAA0312 [Homo sapiens]	29	5.9	L
gi|22090626|dbj|BAC06833.1| HECT domain protein LASU1 [Homo...	29	6.3	L
gi|22538476|ref|NP_113584.2| upstream regulatory element bi...	29	6.3	L
gi|21730336|pdb|1IW2|A Chain A, X-Ray Structure Of Human Co...	29	6.6	S

FIGURE 4.31. *Result of a blastp nr (human) search using RBP4 as a query and (a) a PAM30 scoring matrix or (b) PAM70. Compare the results from (a) to the default BLOSUM62 matrix (Fig. 4.18); distantly related proteins such as apolipoprotein D receive worse scores with the PAM30 matrix. The PAM70 matrix in (b) is designed to detect more distant matches from the database, and several proteins appear that were not apparent using BLOSUM62 or PAM30 matrices.*

```
                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

gi|17433367|ref|XP_066998.1|  similar to odorant-binding pro...    36   0.024
gi|7657407|ref|NP_055396.1|  odorant-binding protein 2B [Hom...    35   0.042
gi|14585875|gb|AAK67639.1|  proteoglycan link protein [Homo ...    33   0.15
gi|20177973|sp|Q8WX39|MUPL_HUMAN  Putative MUP-like lipocali...    31   0.79
gi|17483447|ref|XP_065151.1|  similar to cytochrome P450, su...    30   1.6

                                    Alignments

>gi|17433367|ref|XP_066998.1|  similar to odorant-binding protein [Elephas maximus] [Homo sapiens]
          Length = 103

 Score = 36.2 bits (82), Expect = 0.024
 Identities = 18/58 (31%), Positives = 32/58 (55%)

Query: 27  EVNGDWRTLYIVADNVEKVAEGGSLRAYFQHMECGDECQELKIIFNVKLDSECQTHTV 84
           +V GD  T+++  D+VEK+ E G + +Y   +EC +  + L   F +K +S   T  +
Sbjct: 46  KVAGDRYTIFMAVDDVEKIQENGPITSYIHWIECYNGGETLIFTFYMKDNSAEHTDDI 103

>gi|7657407|ref|NP_055396.1|  odorant-binding protein 2B [Homo sapiens]
 gi|20139234|sp|Q9NPH6|OBP2B_HUMAN  Odorant-binding protein 2b precursor (OBPIIb)
 gi|6900075|emb|CAB71322.1|  odorant binding protein b [Homo sapiens]
 gi|6900077|emb|CAB71323.1|  putative odorant-binding protein b-a [Homo sapiens]
          Length = 170

 Score = 35.4 bits (80), Expect = 0.042
 Identities = 42/185 (22%), Positives = 71/185 (38%), Gaps = 35/185 (18%)

Query: 2   VKFLLIVLALGVSCAHHENLDISPSEVNGDWRTLYIVADN-------VEKVAE------- 47
           +K L + + LG++ A   +  +   ++ G W    +V D            KV+
Sbjct: 1   MKTLFLGVTLGLAAAL--SFTLEEEDITGTWYVKAMVVDKDFPEDRRPRKVSPVKVTALG 58

Query: 48  GGSLRAYFQHMECGDECQELKIIFNVKLDSECQTHTVVGQKHEDGRYTTDYSGRN--YFH 105
           GG L A F  M   D C + KI+                +  E G+Y+  Y GR   Y
Sbjct: 59  GGKLEATFTFMR-EDRCIQKKILMR--------------KTEEPGKYSA-YGGRKLMYLQ 102

Query: 106 VLKKTDDIIFFHNVNVDESGRRQCDLVAGKREDLNKAQKQELRKLAEEYNIPNENTQHLV 165
            L + D  IF+   +    G      + G+  D N+   +E +KL +   +  E+    +
Sbjct: 103 ELPRRDHYIFYCK-DQHHGGLLHMGKLVGRNSDTNREALEEFKKLVQRKGLSEEDIFTPL 161

Query: 166 PTDTC 170
           PT +C
Sbjct: 162 QTGSC 166
```

FIGURE 4.32. *Result of a blastp search of human proteins using rat odorant-binding protein as a query (P08937).*

An important issue in BLAST searching is deciding whether an alignment is significant. Each potential BLAST match should be compared to the query sequence to evaluate whether it is reasonable from both a statistical and a biological point of view. It is more likely that two proteins are homologous if they share similar domain architecture (i.e., motifs or domains; Chapter 8).

WEB RESOURCES

The main website for BLAST searching is that of the National Center for Biotechnology Information (▶http://www.ncbi.nlm.nih.gov/blast/). Within this site are links to the main programs (blastn, blastp, blastx, tblastn, and tblastx). There are many other specialized BLAST websites that will be discussed in Chapter 5.

As an alternative to BLAST searching, one may query other databases such as Pfam and Ensembl. These database search tools will also be discussed in Chapter 5.

An important web resource is the set of BLAST tutorials, courses, and references available at the NCBI BLAST site.

DISCUSSION QUESTIONS

[4-1] Why doesn't anyone offer "Basic Global Alignment Search Tool" (BGAST) to complement BLAST? Would BGAST be a useful tool? What computational difficulties might there be in setting it up?

[4-2] Should you consider a significant expect value to be 1, 0.05, or 10^{-5}? Does this depend on the particular search you are doing?

PROBLEMS

[4-1] The NCBI BLAST page offers a guide for performing specialized searches (http://www.ncbi.nlm.nih.gov/BLAST/; see Fig. 4.33). First guess what the suggested values are for the expect value, word size, and low-complexity filtering. Then check your answers by going to the NCBI website and clicking the hyperlinks for short nucleotide and short protein queries.

[4-2] Why is it that database programs such as BLAST must make a trade-off between sensitivity and selectivity? How does the blastp algorithm address this issue?

[4-3] Protein searches are usually more informative than DNA searches. Do a blastp search using RBP4 (NP_006735), restricting the output to Arthropoda (insects). Next, do a blastn search using the RBP4 nucleotide sequence (NM_006744; select only the nucleotides corresponding to the coding region of the DNA). Which search is more informative? How many databases matches have an *E* value less than 1.0 in each search?

[4-4] The largest gene family in humans is said to be the olfactory receptor family. Do a BLAST search to evaluate how large the family is. *Hint:* As one strategy, first go to LocusLink and enter "olfactory receptor" limiting the organism to *Homo sapiens*. There are over 1100 entries, but this does not tell you whether they are related to each other. Select one accession number and perform a blastp search restricting the organism to human.

[4-5] For the search you just performed in problem 4.1, what happens if you use a scoring matrix that is more suited to finding distantly related proteins?

[4-6] The odorant-binding protein (OBP) was described in rat and cow as a lipocalin that binds odorants selectively. Find the accession number of the rat protein (*hint:* try Entrez proteins) and search for the closest human homolog. What is it?

[4-7] Is the pol protein of HIV-1 more closely related to the pol protein of HIV-2 or to the pol protein of simian immunodeficiency virus (SIV)? Use the blastp program to decide.

[4-8] "The Iceman" is a man who lived 5300 years ago and whose body was recovered from the Italian Alps in 1991. Some fungal material was recovered from his clothing and sequenced. To what modern species is the fungal DNA most related?

[4-9] You perform a BLAST search and a result has an *E* value of about 6×10^{-29}. What does this *E* value mean? What are some parameters on which an *E* value depends?

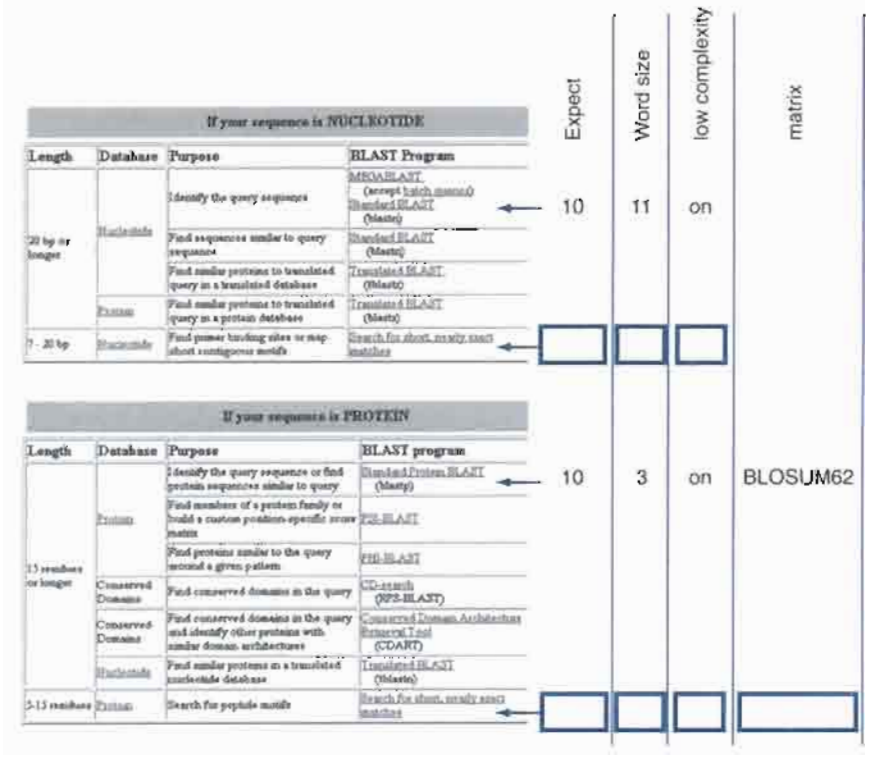


FIGURE 4.33. The NCBI BLAST site includes a guide for a variety of searches with nucleotide or protein queries. For short queries, the expect value and the word size are changed. For protein searches, the suggested scoring matrix is also modified. Guess what the modified values are, then check the BLAST guide (http://www.ncbi.nlm.nih.gov/BLAST).

SELF-TEST QUIZ

[4-1] You have a short DNA sequence. Basically, how many proteins can it *potentially* encode?
(a) 1
(b) 2
(c) 3
(d) 6

[4-2] You have a DNA sequence. You want to know which protein in the main protein database ("nr," the nonredundant database) is most similar to some protein encoded by your DNA. Which program should you use?
(a) blastn
(b) blastp
(c) blastx
(d) tblastn
(e) tblastx

[4-3] Which output from a BLAST search provides an estimate of the number of false positives from a BLAST search?
(a) *E* value
(b) Bit score
(c) Percent identity
(d) Percent positives

[4-4] Match up the following BLAST search programs with their correct descriptions:
blastp__ (a) Nucleotide query against a nucleotide sequence database
blastn__ (b) Protein query against a translated nucleotide sequence database
blastx__ (c) Translated nucleotide query against a protein database
tblastn__ (d) Protein query against a protein database
tblastx__ (e) Translated nucleotide query against a translated nucleotide database

[4-5] Changing which of the following BLAST parameters would tend to yield fewer search results?
(a) Turning off the low-complexity filter
(b) Changing the expect value from 1 to 10
(c) Raising the threshold value
(d) Changing the scoring matrix from PAM30 to PAM70

[4-6] You can limit a BLAST search using any Entrez term. For example, you can limit the results to those containing a researcher's name.
(a) True
(b) False

[4-7] An extreme value distribution:
(a) Describes the distribution of scores from a query against a database
(b) Has a larger total area than a normal distribution
(c) Is symmetric
(d) Has a shape that is described by two constants: μ (mu, the mean) and λ (a decay constant)

[4-8] As the *E* value of a BLAST search becomes smaller:
(a) The value *K* also becomes smaller
(b) The score tends to be larger
(c) The probability *p* tends to be larger
(d) The extreme value distribution becomes less skewed

[4-9] The BLAST algorithm compiles a list of "words" typically of three amino acids (for a protein search). Words at or above a threshold value *T* are defined as:
(a) "Hits" and are used to scan a database for exact matches that may then be extended
(b) Hits and are used to scan a database for exact or partial matches that may then be extended
(c) Hits and are aligned to each other
(d) Hits and are reported as raw scores

[4-10] Normalized BLAST scores (also called bit scores):
(a) Are unitless
(b) Are not related to the scoring matrix that is used
(c) Can be compared between different BLAST searches, even if different scoring matrices are used
(d) Can be compared between different BLAST searches, but only if the same scoring matrices are used

SUGGESTED READING

BLAST searching was introduced in a classic paper by Stephen Altschul, David Lipman, and colleagues (1990). This paper describes the theoretical basis for BLAST searching and describes basic issues of BLAST performance, including sensitivity (accuracy) and speed. Fundamental modifications to the original BLAST algorithm were later introduced, including the introduction of gapped BLAST (Altschul et al., 1997). This paper includes a discussion of specialized position-specific scoring matrices that we will consider in Chapter 5.

William Pearson (1996) provides an excellent description of database searching in an article entitled "Effective Protein Sequence Comparison." Altschul et al. (1994) provide a highly recommended article, "Issues in Searching Molecular Sequence Databases." Marco Pagni and C. Victor Jongeneel (2001) of the Swiss Institute of Bioinformatics provide a technical overview of sequence alignment statistics. This article includes sections on the extreme value distribution, the use of random sequences, local alignment with and without gaps, and BLAST statistics. Another excellent review of alignment statistics was written by Stephen Altschul and Warren Gish (1996).

REFERENCES

Altschul, S. F., Boguski, M. S., Gish, W., & Wootton, J. C. Issues in searching molecular sequence databases. *Nat. Genet.* **6,** 119–129 (1994).

Altschul, S. F., and Gish, W. Local alignment statistics. *Methods Enzymol.* **266,** 460–480 (1996).

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410. (1990).

Altschul, S. F., et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).

Brenner, S. E. Practical database searching. *Bioinformatics: A Trends Guide* **1998,** 9–12 (1998).

Chambaud, I., Heilig, R., Ferris, S., Barbe, V., Samson, D., Galisson, F., Moszer, I., Dybvig, K., Wroblewski, H., Viari, A., Rocha, E. P., Blanchard, A. The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Res.* **29,** 2145–2153 (2001).

Dixon, R. A., et al. Cloning of the gene and cDNA for mammalian beta-adrenergic receptor and homology with rhodopsin. *Nature* **321,** 75–79 (1986).

Downward, J., et al. Close similarity of epidermal growth factor receptor and v-*erb-B* oncogene protein sequences. *Nature* **307,** 521–527 (1984).

Ermolaeva, M. D., White, O., and Salzberg, S. L. Prediction of operons in microbial genomes. *Nucleic Acids Res.* **29,** 1216–1221 (2001).

Ferretti, J. J., et al. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc. Natl. Acad. Sci. USA* **98,** 4658–4663 (2001).

Gish, W., and States, D. J. Identification of protein coding regions by database similarity search. *Nat. Genet.* **3,** 266–272 (1993).

Gotoh, O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.* **264,** 823–838 (1996).

Gumbel, E. J. *Statistics of Extremes.* Columbia University Press, New York, 1958.

International Human Genome sequencing consortium. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921 (2001).

Karlin, S., and Altschul, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* **87,** 2264–2268 (1990).

Karlsson, H., et al. Retroviral RNA identified in the cerebrospinal fluids and brains of individuals with schizophrenia. *Proc. Natl. Acad. Sci. USA* **98,** 4634–4639 (2001).

Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453 (1970).

Pagni, M., and Jongeneel, C. V. Making sense of score statistics for sequence alignments. *Brief. Bioinform.* **2,** 51–67 (2001).

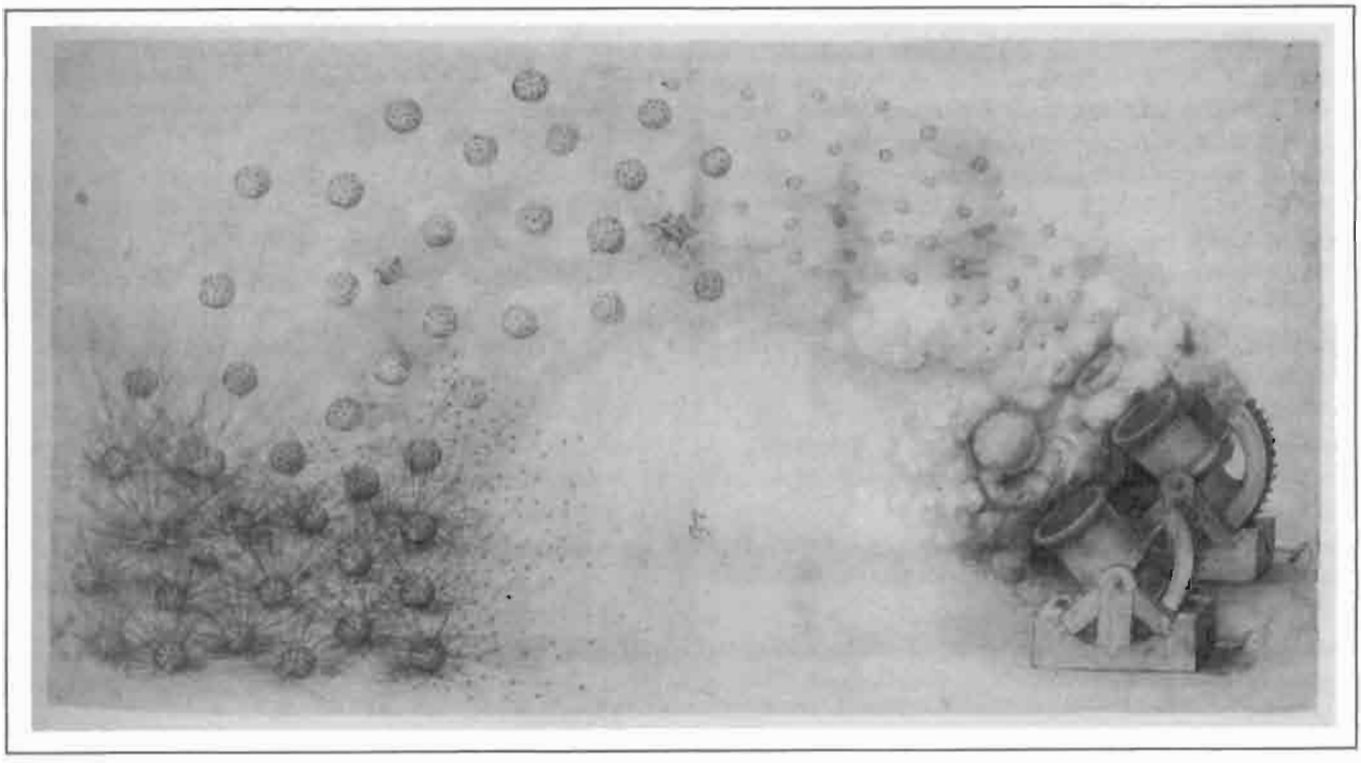
Park, J., et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284,** 1201–1210 (1998).

Pearson, W. R. Effective protein sequence comparison. *Methods Enzymol.* **266,** 227–258 (1996).

Smith, T. F., and Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).

Tettelin, H., et al. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293,** 498–506 (2001).

Wootton, J. C., and Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266,** 554–571 (1996).



Leonardo da Vinci designed mortars with explosive shells (from the Codex Atlanticus, 33 recto). The balls are filled with holes to maximize the deadly force of the shrapnel they contain. The drawing dates from c.1485.

5

Advanced BLAST Searching

INTRODUCTION

In Chapters 3 and 4 we introduced pairwise alignments and BLAST searching. BLAST searching allows one to search a database to find what proteins or genes are present. BLAST searches can be very versatile, and in this chapter we will cover several advanced BLAST searching techniques.

We begin with a brief overview of the kinds of specialized BLAST resources that are available to help solve many kinds of research questions. Why do we need specialized BLAST sites? The answer is that we often have very specialized research problems, and the separate BLAST sites are enormously helpful in our bioinformatic approaches. Common questions we might want to ask with BLAST searching are: Is there a gene that is related to my favorite gene in a particular organism? What motifs are present in my protein? Are there microbial homologs of my gene? We will also address the problem of how to discover novel genes using BLAST searches.

SPECIALIZED BLAST SITES

So far, we have used two BLAST resources, both from the NCBI website: BLAST 2 Sequences and the standard five BLAST programs. In this chapter we will describe other, specialized BLAST (and BLAST-related) programs. The general categories of available BLAST sites can be organized as follows: organism-specific BLAST sites, BLAST sites that allow searches of specific molecules, and specialized database

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

search algorithms. Several of these specialized algorithms greatly extend the sensitivity of BLAST searching.

Organism-Specific BLAST Sites

We have seen that for standard BLAST searches at the NCBI website the output can be restricted to a particular organism. There are many other entire databases that consist of molecular sequence data from a specific organism (Table 5.1). Often, the data include unfinished sequences that have not yet been deposited in GenBank. Thus, searches of these databases can yield useful information. If you have a protein or DNA sequence with no apparent matches in standard NCBI BLAST searches, then searching these specialized databases provides a more exhaustive search. Also, as described below, some of these databases present unique output formats and/or search algorithms.

TABLE 5-1 Organism-Specific BLAST and FASTA Sites

These sites provide access to databases with sequences that are not available in standard NCBI BLAST searches. Some use the WU BLAST 2.0 algorithm (described below)

Organism	Description	URL
Bacteria	Cyanobase (Cyanobacteria)	▸http://www.kazusa.or.jp/cyano/search.html
Streptomyces coelicolor	The Wellcome Trust Sanger Institute	▸http://www.sanger.ac.uk/Projects/S_coelicolor/blast_server.shtml
Dictyostelium discoideum	Dicty blast	▸http://glamdring.ucsd.edu/others/dsmith/dictydb.html
	D. discoideum Genome Project	▸http://genome.imb-jena.de/dictyostelium/BlastDictyForm.html
Fungi		
Schizosaccharomyces pombe	The Wellcome Trust Sanger Institute	▸http://www.sanger.ac.uk/Projects/S_pombe/blast_server.shtml
Aspergillus fumigatus	The Wellcome Trust Sanger Institute	▸http://www.sanger.ac.uk/Projects/A_fumigatus/
Candida albicans	The Wellcome Trust Sanger Institute	▸http://www.sanger.ac.uk/Projects/C_albicans/Urlbblast_server.shtml
Neurospora crassa	Whitehead Institute	▸http://www-genome.wi.mit.edu/annotation/fungi/neurospora/
Pneumocystis carinii	The Wellcome Trust Sanger Institute	▸http://www.sanger.ac.uk/Projects/P_carinii/blast_server.shtml
Saccharomyces cerevisiae	MIPS	▸http://mips.gsf.de/cgi-bin/blast/blast_page?genus=yeast
White rot	Department of Energy Joint Genome Institute	▸http://www.jgi.doe.gov/programs/whiterot/whiterot_mainpage.html
Insect		
Drosophila melanogaster	FlyBLAST	▸http://www.fruitfly.org/blast/
Mosquito	An NCBI BLAST of culicidae from the Mosquito Genomics WWW server (▸http://klab.agsci.colostate.edu/)	▸http://mosquito.colostate.edu/cgi-bin/ace/blast/MsqDB
Anopheles gambiae	AnoDB	▸http://konops.imbb.forth.gr/cgi-bin/blast2.pl
Microbial Genomes	Allows BLAST of both completed and unfinished genomes	▸http://www.ncbi.nlm.nih.gov/BLAST/
Various	Allows BLAST of both completed and unfinished genomes	▸http://www.tigr.org
	Comprehensive microbial resource at TIGR	▸http://tigrblast.tigr.org/cmr-blast/
	Dozens of organism-specific BLAST searches	▸http://www.sanger.ac.uk/Projects/Microbes/

(*Continued*)

TABLE 5-1 *(Continued)*

Organism	Description	URL
Parasites		
Various	At EBI	http://www.ebi.ac.uk/blast2/parasites.html
	From the Oswaldo Cruz Institute, Brazil	http://www.dbbm.fiocruz.br/genome/parasite-genome/parasite_blast_server.html
Phytophthora	A plant pathogen studied at http://www.ncgr.org	http://www.ncgr.org/pgc/blast/blast.html
Plants: *Arabidopsis thaliana*	Thale cress	http://genome-www2.stanford.edu/cgi-bin/AtDB/nph-blast2atdb
Protozoa		
Babesia bovis	The Wellcome Trust Sanger Institute	http://www.sanger.ac.uk/Projects/B_bovis/blast_server.shtml
Giardia lamblia	Marine Biological Laboratory, Woods Hole	http://www.mbl.edu/Giardia/blast.html
Plasmodium chabaudi	The Wellcome Trust Sanger Institute	http://www.sanger.ac.uk/Projects/P_chabaudi/blast_server.shtml
Plasmodium falciparum	This is the organism that causes malaria	http://www.ncbi.nlm.nih.gov/BLAST/
	Plasmodium genome resource	http://plasmodb.org/
	The Wellcome Trust Sanger Institute	http://www.sanger.ac.uk/Projects/P_falciparum/blast_server.shtml
Tetrapygus niger	Sea urchin BLAST	http://sea-urchin.caltech.edu/genome/sequence/blast.py
Various organisms	TIGR BLAST	http://www.tigr.org
	OnmiBLAST server at The Wellcome Trust Sanger Institute	http://www.sanger.ac.uk/DataSearch/omniblast.shtml
	The Gene2EST BLAST server	http://woody.embl-heidelberg.de/gene2est/
	The oral BLAST server	http://199.93.248.241/blast/blast.html
Vertebrate		
Ciona intestinalis	Sea squirt	http://www.jgi.doe.gov/programs/ciona/ciona_mainpage.html
Danio rerio	Zebrafish at The Wellcome Trust Sanger Institute	http://www.sanger.ac.uk/Projects/D_rerio/blast_server.shtml
Fugu	Includes *Fugu* (the pufferfish) and zebrafish	http://fugu.hgmp.mrc.ac.uk/blast/
	Department of Energy Joint Genome Institute	http://www.jgi.doe.gov/programs/fugu/fugu_mainpage.html
Gallus gallus	From ChickMap (http://www.ri.bbsrc.ac.uk/chickmap/ChickMapHomePage.html)	http://www.ri.bbsrc.ac.uk/cgi-bin/est-blast/blast.pl
Homo sapiens	Allows searches against specific chromosomes	http://www.ncbi.nlm.nih.gov/BLAST/
	Premiere human genome site	http://www.Ensembl.org
Mus musculus	Chromosome-specific mouse BLAST	http://www.ncbi.nlm.nih.gov/genome/seq/MmBlast.html
	Mouse Genome Informatics at Jackson Laboratories	http://mouseblast.informatics.jax.org/prototype/
	The Wellcome Trust Sanger Institute and Ensembl	http://mouse.Ensembl.org/
	The Wellcome Trust Sanger Institute	http://mouse.Ensembl.org/perl/blastview
	National Institute of Aging	http://lgsun.grc.nia.nih.gov/cDNA/cDNA.html
Xenopus	Contains data from both *X. laevis* and *X. tropicalis*	http://www.sanger.ac.uk/Projects/X_tropicalis/blast_server.shtml

(Continued)

TABLE 5-1 *(Continued)*

Organism	Description	URL
Worms		
Brugia malayi	The Wellcome Trust Sanger Institute	▸http://www.sanger.ac.uk/Projects/B_malayi/blast_server.shtml
	B. Malayi BLAST server	▸http://www.dbbm.fiocruz.br/genome/parasite-genome/Brugia_blast_server.html
Caenorhabditis elegans	WormBLAST; also other organisms; Genome Sequencing Center, Washington University	▸http://genome.wustl.edu/gsc/Blast/client.pl
C. elegans	The Wellcome Trust Sanger Institute	▸http://www.sanger.ac.uk/Projects/C_elegans/blast_server.shtml
C. elegans	Steve Jones' Genome Project DNA/protein databases	▸http://www.bio.ic.ac.uk/fgn/blastform/elegansblast.html

Ensembl BLAST

The Wellcome Trust Sanger Institute website is ▸http://www.sanger.ac.uk/. The EBI is at ▸http://www.ebi.ac.uk/. Ensembl's human BLAST server is at ▸http://www.ensembl.org/perl/blastview, and Ensembl BLAST servers for mouse and other organisms can be found through ▸http://www.ensembl.org/. We will discuss "finished" sequences in Chapters 12 and 17.

Project Ensembl is a joint effort of The Wellcome Trust Sanger Institute and the European Bioinformatics Institute (EBI). The Ensembl website provides a comprehensive resource for studying the human genome and other genomes (see Chapters 16 and 17). The Ensembl BLAST server allows the user to search the Ensembl database, including the most finished sequence. As an example, paste in the plain-text amino acid sequence of human RBP4 (accession NP_006735) and perform a tblastn search (Fig. 5.1). One significant alignment is returned (with a P value of 3.8×10^{-81}; AL356214.00005). Following this link leads to a summary page with dozens of fields of information, including the chromosome band (10q23.33), neighboring genes, protein and DNA database links (including NCBI links), polymorphisms, mouse homologies, and expression data. The output also consists of a graphical output showing the location of the database matches by chromosome (Fig. 5.2).

The Institute for Genomic Research (TIGR) BLAST

TIGR is at ▸http://www.tigr.org, and the TIGR Gene Indices host a BLAST server ▸(http://tigrblast.▸tigr.org/tgi/).

TIGR is an institute that has sequenced the complete genomes of many organisms, and TIGR scientists have sequenced parts of various genomes, including human, rice, thale cress, potato, parasites, microbes, and mouse. TIGR BLAST searches feature access to many unique databases, and the searches allow the user to choose from almost 80 substitution matrices. From the TIGR website, select TIGR Gene Indices, then BLAST search. Enter the human RBP4 protein sequence (obtained from Entrez record NP_006735) into the text box (Fig. 5.3). Note that accession numbers may not be used, so copy and paste the protein sequence in FASTA format. Choose the appropriate algorithm (tblastn for a protein query), and select EGO (Eukaryotic Gene Orthologs). Hits are returned ranked by probability scores (note that P values rather than E values are used) (Fig. 5.4). The matches are to entries with TC ("tentative consensus") or THC ("tentative human consensus"). Follow the link for one of these and you will find the DNA sequence for that entry, a description of the corresponding ESTs, an expression summary describing the regional localization of the cDNA libraries from which ESTs encoding this protein were isolated, and a table of RBP orthologs from several species. Additionally, there is a link to tentative ortholog groups (TOGs) (Fig. 5.5). These groups provide a graphical overview of known orthologs within the TIGR databases, including a table of pairwise alignments (partially shown in Fig. 5.5) and a multiple sequence alignment of the DNA sequences within this group.

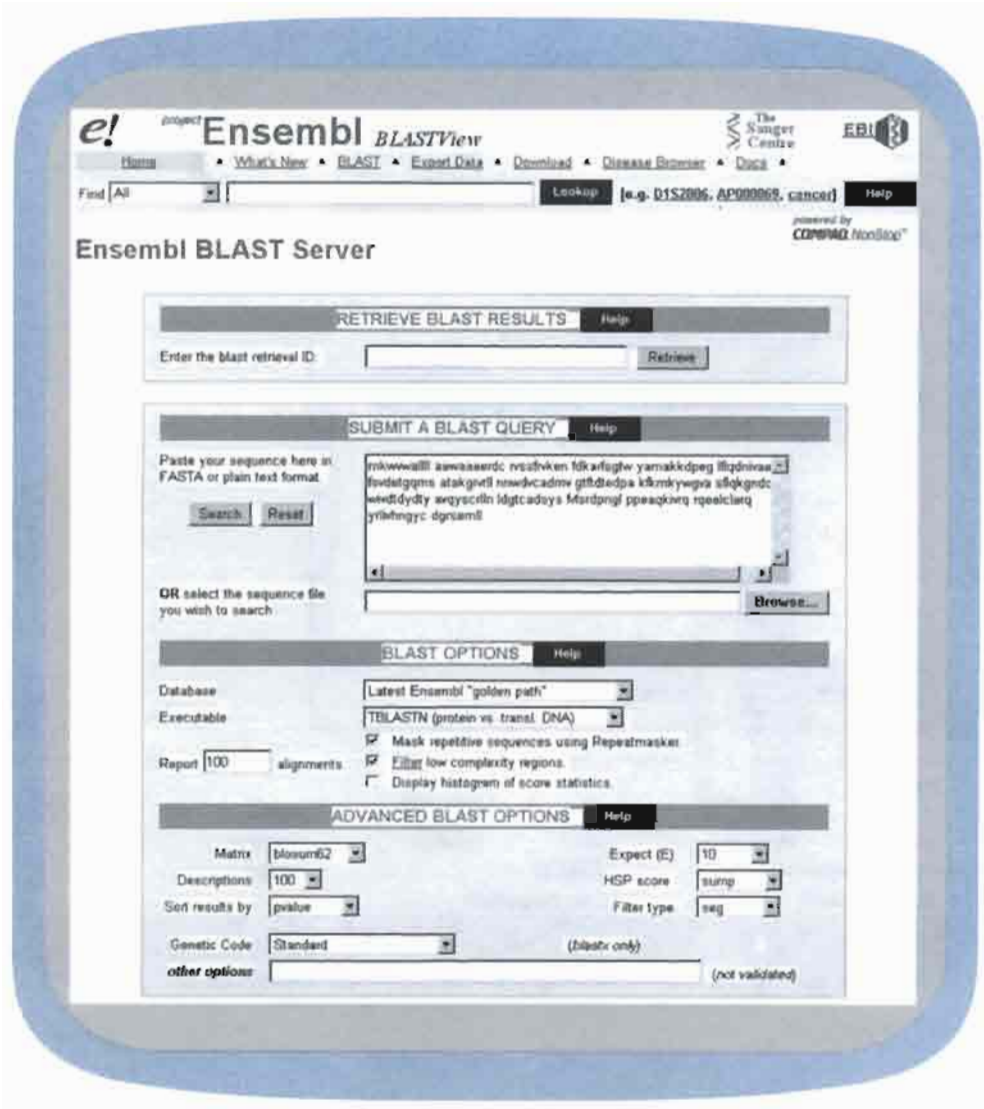


FIGURE 5.1. Ensembl BLAST server (http://www.ensembl.org/perl/blastview) allows queries of a human genome database.

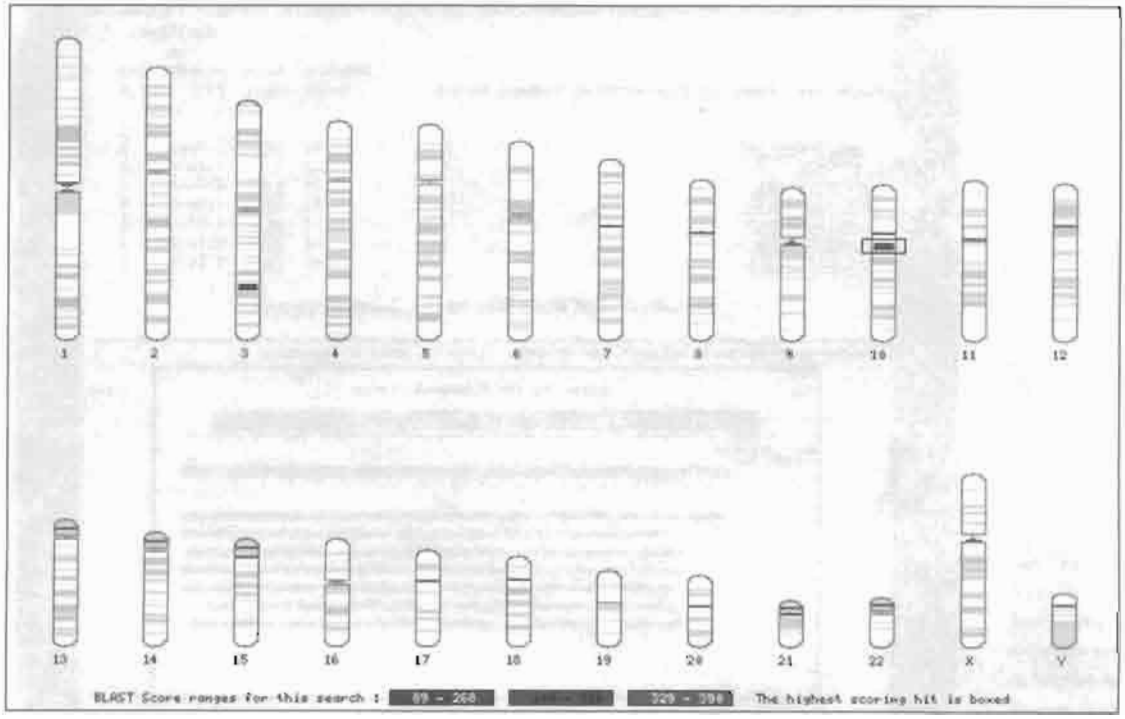


FIGURE 5.2. Output of a BLAST search of the Ensembl database using RBP4 as a query. The results are presented in a graphical format by chromosome, showing the best match to the long arm of chromosome 10 near the centromere. Weaker matches to paralogs on other chromosomes are also evident.

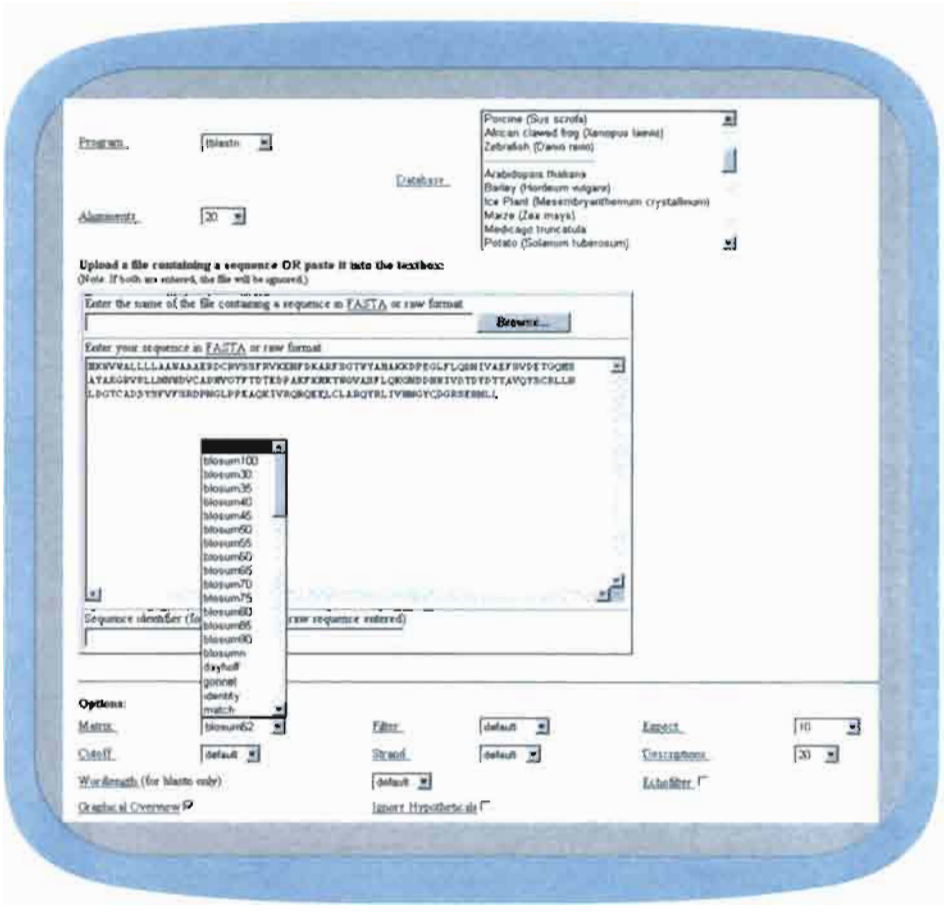


FIGURE 5.3. BLAST search from TIGR allows the choice of databases from various organisms as well as optional parameters such as a choice from dozens of substitution matrices. Note that the input sequence should be in the FASTA format.

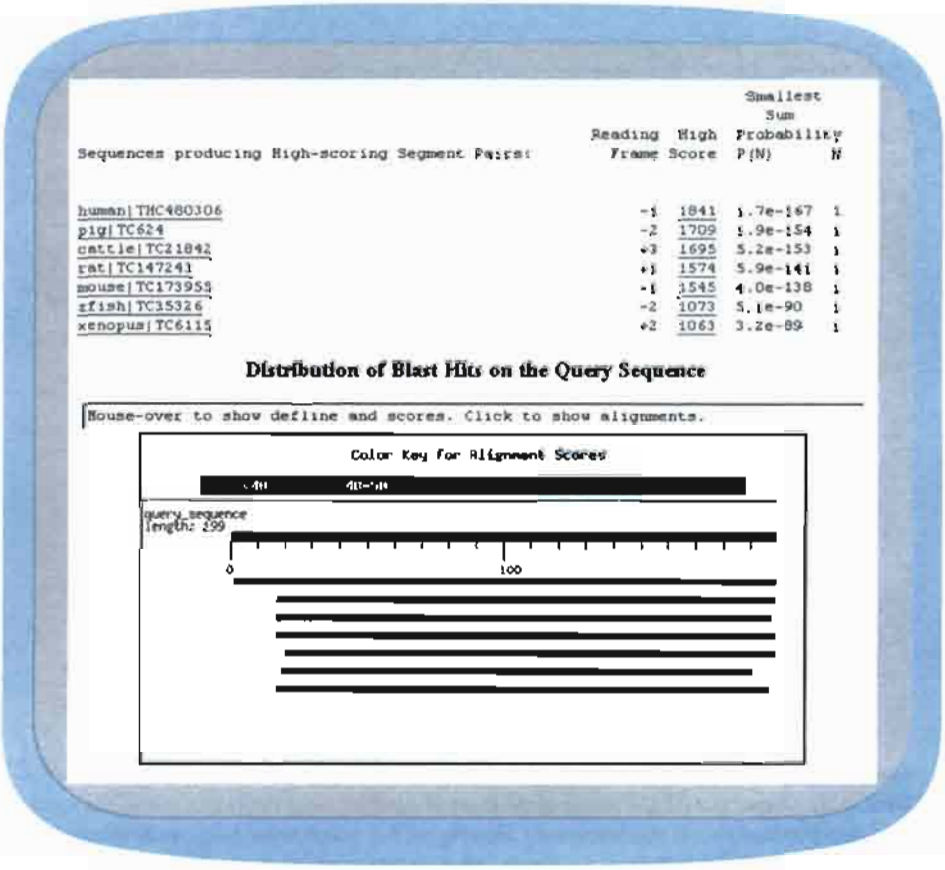


FIGURE 5.4. The TIGR BLAST output resembles that of NCBI BLAST. Note that fewer organisms are returned. Also, there is typically only one entry per species because redundant or partial sequences from assorted databases are unified into one accession number. Abbreviations: TC, tentative consensus; THC, tentative human consensus.

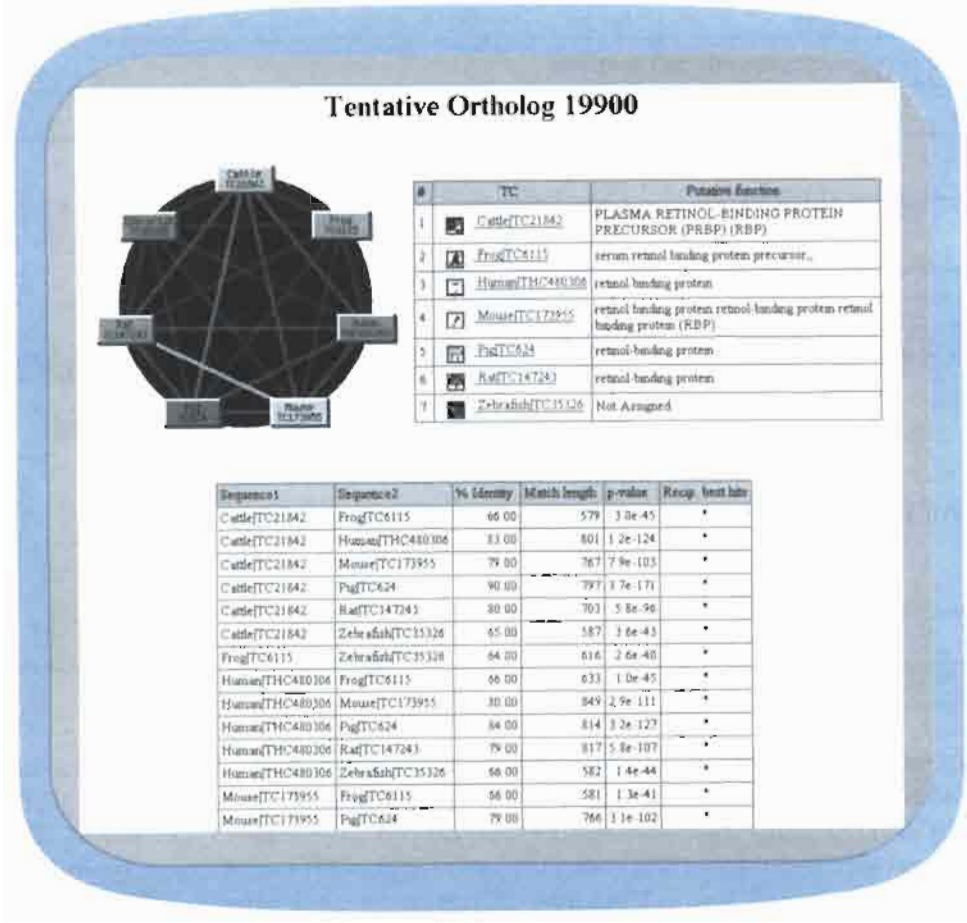


FIGURE 5.5. BLAST output from a TIGR search may include links to tentative ortholog groups (TOGs) such as this one for RBPs. This TOG also includes a multiple sequence alignment of the DNA sequences (not shown).

The features offered in Ensembl and TIGR BLAST searches typify the ways that many websites allow database searches that offer different choices for query inputs, choices of search algorithms, and links to many kinds of information as part of the search output.

Molecule-Specific BLAST Sites

Several BLAST sites are specialized to allow searches of particular molecules (Table 5.2). As an example, the NCBI BLAST page includes IgBLAST for the analysis of immunoglobulins. IgBLAST reports the three germline V genes, two D and two J genes that show the closest match to the query sequence. It also annotates the immunoglobulin domains. Finally, IgBLAST matches the BLAST search results to the closest germline V genes.

Some servers specialize in searches of proteins (rather than nucleotides). The ProDom BLAST server in Toulouse provides a standard input for a BLAST search and an output that resembles an NCBI search but also includes a graphical view of the protein domain structure (Fig. 5.6).

Specialized BLAST Servers and BLAST-Related Algorithms

We have focused on the standard BLAST algorithms at NCBI, but many other algorithms are available. Table 5.3 lists some of these, such as the WU BLAST 2.0

You can learn more about WU BLAST 2.0 at ▸http://blast.wustl.edu.

TABLE 5-2 Molecule-Specific BLAST and BLAST-Related Sites

These sites provide databases of particular molecules for selected searches

Molecule	Description	URL
Immunoglobulins	IgBLAST	▸ http://www.ncbi.nlm.nih.gov/BLAST/
MS BLAST	Mass spectrometry driven BLAST (WU BLAST 2.0)	▸ http://dove.embl-heidelberg.de/Blast2/msblast.html
Proteins	Network Protein Sequence Analysis at Pole BioInformatique Lyonnais	▸ http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_blast.html
	EMBnet-CH BLAST Network Service	▸ http://kr.expasy.org/cgi-bin/BLASTEMBnet-CH.pl
	ProDom BLAST with graphical output (Toulouse)	▸ http://prodes.toulouse.inra.fr/prodom/doc/blast_form.html
Single-nucleotide polymorphisms (SNPs)	At NCBI	▸ http://www.ncbi.nlm.nih.gov/SNP/snpblastByChr.html
Transcript assembly program	EST-based gene finder (WU BLAST 2.0)	▸ http://sapiens.wustl.edu/~zkan/TAP/
Vectors (VecScreen)	Useful to detect vector contamination of sequences	▸ http://www.ncbi.nlm.nih.gov/BLAST/

algorithm. Developed at Washington University, WU BLAST 2.0 is related to the traditional NCBI BLAST algorithms, as both were developed from the original NCBI BLAST algorithms that did not permit gapped alignments. WU BLAST 2.0 may provide faster speed and increased sensitivity, and it includes a variety of options such as a full Smith–Waterman alignment on some pairwise alignments of database matches. WU BLAST 2.0 runs on a variety of computer servers, including EMBL/EBI (Fig. 5.7) and TIGR (Fig. 5.3). Other servers use the BLAST algorithm, typically with minor modifications (see, e.g., the DNA Database of Japan; Fig. 5.8). The FASTA algorithm is also widely available (Table 5.3) (Pearson and Lipman 1988).

Other database searching algorithms such as the BLAST-like alignment tool (BLAT), PSI-BLAST, and Pfam use search strategies that are fundamentally different than BLAST. We will examine some of these in greater detail.

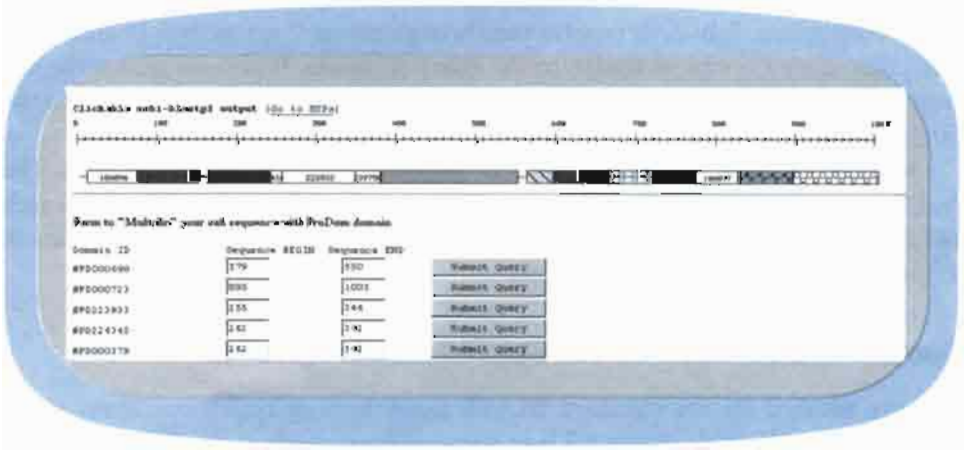


FIGURE 5.6. BLAST output of the ProDom server ▸ (http://prodes.toulouse.inra.fr/prodom/2002.1/html/form.php) provides graphical output of protein domains. Here, the query is HIV-1 pol (NP_057849). There are further links to information on the individual domains of HIV-1.

TABLE 5-3 Specialized BLAST Algorithms and Distinct BLAST-Related Algorithms Available on World Wide Web

Algorithm	Description	URL
PSI-BLAST	Position-specific iterated BLAST	NCBI
PHI-BLAST	Pattern-hit-initiated BLAST	NCBI
MegaBLAST	For large DNA queries	NCBI
RPS-BLAST	See Chapter 6	NCBI
NCBI BLAST2	EMBL/European Bioinformatics Institute	▸ http://www2.ebi.ac.uk/blastall/
FASTA3	EMBL/European Bioinformatics Institute	▸ http://www2.ebi.ac.uk/fasta3/
WU BLAST 2.0	EMBL/European Bioinformatics Institute	▸ http://www2.ebi.ac.uk/blast2/
WU BLAST 2.0	EMBL	▸ http://dove.embl-heidelberg.de/Blast2/
WU BLAST 2.0	Genome Sequencing Center, Washington University	▸ http://genome.wustl.edu/gsc/Blast/client.pl
WU BLAST 2.0	Institut Pasteur	▸ http://bioweb.pasteur.fr/seqanal/interfaces/blastwu.html
WU BLAST 2.0	Mouse Genome Informatics at Jackson Labs	▸ http://mouseblast.informatics.jax.org/prototype/
WU BLAST 2.0	Transcript assembly program	▸ http://sapiens.wustl.edu/~zkan/TAP/
WU BLAST 2.0	Mendel Bioinformatics Group	▸ http://www.mendel.ac.uk/blastaccess.html
WU BLAST 2.0	The Institute for Genomic Research	▸ http://tigrblast.tigr.org/tgi/
BLAST	Swiss Institute of Bioinformatics and the Swiss Center for Scientific Computing	▸ http://www.ch.embnet.org/software/BottomBLAST.html?
FASTA	DNA Database of Japan	DDBJ
BLAST	DNA Database of Japan	DDBJ
SSEARCH	DNA Database of Japan	DDBJ
PSI-BLAST	DNA Database of Japan	DDBJ
FASTA	From the University of Virginia (Bill Pearson)	▸ http://alpha10.bioch.virginia.edu/fasta/
SSEARCH	From the University of Virginia (Bill Pearson)	▸ http://alpha10.bioch.virginia.edu/fasta/
BSU	BLAST search updater	▸ http://athena.bioc.uvic.ca/genomes/index.html#BSU
MSPcrunch	Used for analysis of BLAST results	▸ http://www.embnet.se/new/mspup.html
Sim4	Program to align cDNA and genomic DNA	▸ http://pbil.univ-lyon1.fr/sim4.html
Sim4	Downloadable software from Penn State Bioinformatics Group	▸ http://bio.cse.psu.edu/
BLAT	BLAST-like tool to align cDNA and genomic DNA	▸ http://genome.ucsc.edu

Abbreviations: NCBI refers to▸ http://www.ncbi.nlm.nih.gov/BLAST/. DDBJ refers to search tools at the DNA Database of Japan▸ (http://www.ddbj.nig.ac.jp/E-mail/homology.html).

BLAST-Like Alignment Tools to Search Genomic DNA Rapidly

As genomic DNA databases grow in size (Chapters 12–17), it becomes increasingly common to search them using protein queries or DNA sequences corresponding to expressed transcripts as queries. This is a specialized problem:

1. The genomic DNA includes both exons (regions corresponding to the coding sequence) and introns (intervening, noncoding regions of genes). Ideally, an alignment tool should find the exons in genomic DNA.
2. Genomic DNA often has sequencing errors that should be taken into account.

We will discuss exons and introns in Chapters 6 (on gene expression), 12 (on genomes), and 16 and 17 (on eukaryotic genomes).

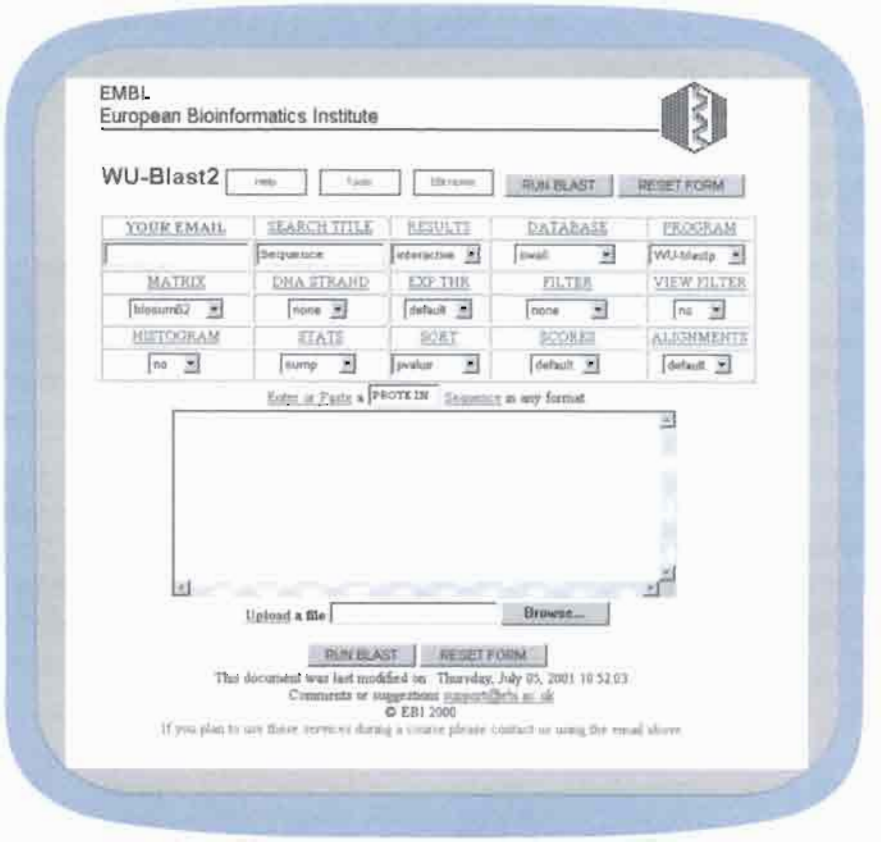


FIGURE 5.7. The European Bioinformatics Institute (EBI; ▸http://www.ebi.ac.uk/) offers a WU-BLAST2 site ▸(http://www2.ebi.ac.uk/blast2/).

3. In some cases, researchers want to compare genomic DNA between closely related organisms, such as human and mouse.
4. Algorithms are needed to find small differences between DNA sequences, such as single-nucleotide polymorphisms (SNPs; Chapter 18).

Several BLAST-like algorithms have been written to address these needs:

1. The NCBI BLAST site includes Megablast, a blastn-like algorithm that is optimized for the rapid alignment of very large DNA queries (Zhang et al., 2000).
2. Sim4 uses a BLAST-like algorithm to determine high-scoring segment pairs (HSPs; Chapter 4) and to extend them in both directions (Florea et al., 1998). A dynamic programming algorithm identifies a chain of HSPs that could represent a gene. For example, the program searches for potential splice signals at exon-intron boundaries.
3. Sequence Search and Alignment by Hashing Algorithm, abbreviated SSAHA, is designed to search large DNA databases very rapidly (Ning et al., 2001). The SSAHA converts a DNA database (of up to billions of base pairs) into a hash table, which can then be searched quickly for matches.
4. BLAT is designed to perform extremely rapid genomic DNA searches (Kent, 2002). Like SSAHA, the BLAT algorithm is in some ways a mirror image of BLAST. BLAST parses a query sequence into words and then searches a database with words above a threshold score. Two proximal hits are extended. BLAT parses an entire genomic DNA database into an index of words. These words consist of all nonoverlapping 11-mers in the genome

You can access sim4 through a server at ▸http://pbil.univ-lyon1.fr/sim4.html. Alternatively, you can download the program from a website of the sim4 authors (including Webb Miller at Pennsylvania State University) at ▸http://globin.cse.psu.edu/.

SSAHA is available at the Ensembl web server ▸(http://www.ensembl.org). The SSAHA home page is ▸http://www.sanger.ac.uk/Software/analysis/SSAHA/. A hash table contains data (e.g., a list of words having a length of 14 nucleotides in a DNA database) and associated information (e.g., the positions in genomic DNA of each of those words).

BLAT is accessible on the web at ▸http://genome.ucsc.edu. This is one of the main human genome browsers, and we will explore it in Chapter 17.

FIGURE 5.8. *BLAST page from the DNA Database of Japan website (http://www.ddbj.nig.ac.jp/E-mail/homology.html).*

(excluding repetitive DNA sequences). BLAT then searches a query using words from the database.

BLAT offers a variety of additional features (Kent, 2002):

- While BLAST triggers an extension with two hits, BLAT triggers extensions on multiple strong hits.
- BLAT is designed to find matches between queries that share 95% nucleotide identity or more. While it is in some ways similar to the Megablast, Sim4, and SSAHA programs, it is orders of magnitude faster.
- BLAT searches for intron–exon boundaries, essentially building a model of a gene structure. It only uses each nucleotide derived from an mRNA query once (as is appropriate from a biological perspective), rather than searching only for highest scoring segment pairs.

FINDING DISTANTLY RELATED PROTEINS: POSITION-SPECIFIC ITERATED BLAST (PSI-BLAST)

Many homologous proteins share only limited sequence identity. Such proteins may adopt the same three-dimensional structures (based on methods such as X-ray crystallography), but in pairwise alignments they may have no apparent similarity.

We have seen that scoring matrices are sensitive to protein matches at various evolutionary distances. For example, we compared the PAM250 to the PAM10 log-odds matrices (Figs. 3.14 and 3.15) and saw that the PAM250 matrix provides a superior scoring system for the detection of distantly related proteins. In performing a database search with BLAST, we can adjust the scoring matrix to try to detect distantly related proteins. Even so, many proteins in a database are too distantly related to a query to be detected using a standard blastp search.

Position-specific iterated BLAST (abbreviated PSI-BLAST or ψ-BLAST) is a specialized kind of BLAST search that is often more sensitive than a regular BLAST search (Zhang et al., 1998; Altschul et al., 1997; Schaffer et al., 2001). The purpose of using PSI-BLAST is to look deeper into the database to find distantly related proteins that match your protein of interest. In many cases, when a complete genome is sequenced and the predicted proteins are analyzed to search for homologs, PSI-BLAST is the algorithm of choice.

You can access PSI-BLAST at ▸ http://www.ncbi.nlm.nih.gov/BLAST and at other servers (see Web Resources, Table 5.5).

PSI-BLAST is performed in five steps:

1. A normal blastp search uses a scoring matrix (such as BLOSUM62, the default scoring matrix) to perform pairwise alignments of your query sequence (such as RBP) against the database. PSI-BLAST also begins with a protein query that is searched against a database at the NCBI website.
2. PSI-BLAST constructs a multiple sequence alignment from an initial blastp-like search and then creates a specialized, individualized search matrix (also called a profile) based on that multiple alignment.
3. This position-specific scoring matrix (PSSM) is then used as a query to search the database again. (Your original query is not used.)
4. PSI-BLAST estimates the statistical significance of the database matches, essentially using the parameters we described for gapped alignments.
5. The search process is continued iteratively, typically about five times. At each step a new profile is used as the query. You can stop the search process at any point—whenever few new results are returned or when the program reports "convergence" because no new results are found.

We have seen a multiple sequence alignment from a BLAST output in Figure 4.12, and we will examine this topic in Chapter 10.

PSI-BLAST is conceptually related to reverse position-specific BLAST (RPS-BLAST), which will be described in Chapter 8. Both PSI-BLAST and RPS-BLAST use multiple alignments and PSSMs, but the purpose of RPS-BLAST is to search a protein query against a large set of predefined PSSMs to identify conserved protein domains in the query.

We can illustrate the dramatic results of the PSI-BLAST process as follows. Go to PSI-BLAST at NCBI and enter the protein accession number of human RBP4 (NP_006735). Using the default parameters, perform a search of the nonredundant database. There are 110 hits (as of December 2001; your results will likely vary). Half of these are above the inclusion threshold (set as a default at $E = 0.005$), and by inspection these are all called RBP (from various species) or apolipoprotein D (another lipocalin). There are also over 50 database matches that are worse than the inclusion threshold: These do not have significant E values. Some of these distantly related matches (such as insecticyanins) are authentic lipocalins, based on criteria such as having similar three-dimensional structures and related biological functions as carrier proteins. Other proteins on this list are viral and appear to be true negatives.

You can adjust the inclusion threshold. Try E values of 0.5 or 0.00005 to see what happens to your search results. If you set the E value too low, you will only see very closely related homologs. If you set E too high, you will probably find false-positive matches.

Through this initial step, the PSI-BLAST search is performed in a manner nearly identical to a standard blastp search, using some amino acid substitution matrix such as BLOSUM45. However, the program creates a multiple sequence alignment from the initial database matches. By analyzing this alignment, the PSI-BLAST program then creates a PSSM. The original query sequence serves as a template for this profile.

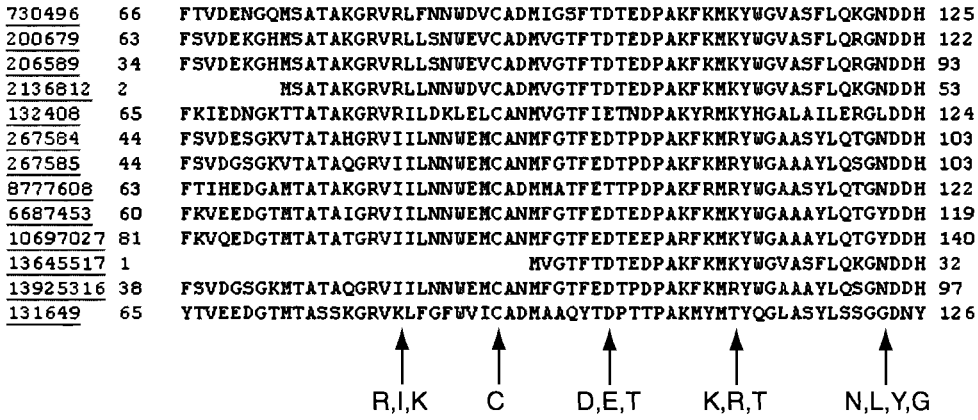


FIGURE 5.9. *PSI-BLAST search begins with a standard blastp-like search. The output is used to generate a profile or PSSM. A PAM or BLOSUM matrix describes the likelihood that one amino acid will be substituted for another, based on a statistical analysis of thousands of proteins. The PSSM is created specifically for the protein query of the PSI-BLAST search. The figure shows a portion of the initial PSI-BLAST output (shown using a query anchored without identities alignment view). The arrows point to examples of columns of amino acids in the alignment and the actual amino acid residues that are tolerated in each position. Some of the positions are invariant (such as C), while other columns show aligned residues (such as R, I, K) that tolerate amino acid substitutions that may be unique to this particular group of proteins.*

Consider a portion of the BLAST output so far viewed as a multiple sequence alignment (Fig. 5.9). In one column, the amino acid residues R, I, and K are found. Substitutions of R and K are quite common in general, but it is rare to substitute I for either of these basic residues. The key idea of PSI-BLAST is that the information at each position of the multiple sequence alignment provides information about the accepted mutations in the query and its nearest database matches. That information forms the basis of a matrix that can be used to search the database with more sensitivity than a standard BLOSUM or PAM matrix can provide.

For a query of length L, PSI-BLAST generates a PSSM of dimension $L \times 20$. The rows of each matrix have a length L equal to the query sequence. A portion of a PSSM derived from a PSI-BLAST search using RBP4 as a query is shown in Figure 5.10. The columns have the 20 common amino acids. Look at the scores given to alanine (at positions 6, 11, 12, 14–16, and 42). In some cases, alanine scores a +4 (this is also the score from a BLOSUM62 matrix, as shown in Fig. 3.17). However, alanine occasionally scores a +2, +3, or +5 at various positions in the query sequence. This highlights the benefit of a scoring matrix that is customized to a query. Next consider the tryptophan at position 40 that is part of the nearly invariant GXW pattern present in hundreds of lipocalins. This W (as well as several other tryptophans) scores a +12 in position 40, but a different W (at position 13) scores only a +7. These examples illustrate one of the main advantages of PSI-BLAST: The PSSM reflects a more accurate estimate of the probabilities with which amino acid substitutions occur at various positions.

In the first iteration of PSI-BLAST, a database search uses this scoring matrix and results in the identification of sequences with values better than a cutoff (e.g., $E = 0.005$). The query sequence forms a template for a multiple sequence alignment. Redundant sequences (having at least 94% amino acid identity in a pairwise alignment of any two sequences in the matrix) are eliminated. This ensures that a group of very closely related sequences will not overly bias the construction of the PSSM. In the original implementation of PSI-BLAST, pairwise alignment

The PSI-BLAST search is not identical to a blastp search at the first step because PSI-BLAST uses composition-based statistics to account for the composition of the query and the database sequences (Schaffer et al., 2001). The scoring matrix may not correspond exactly to BLOSUM62 or other matrices you select, but it will be very closely similar.

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-2	-2	-3	-2	-1	-2	-3	-2	1	2	-2	6	0	-3	-2	-1	-2	-1	1
2	K	-1	1	0	1	-4	2	4	-2	0	-3	-3	3	-2	-4	-1	0	-1	-3	-2	-3
3	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
4	V	0	-3	-3	-4	-1	-3	-3	-4	-4	3	1	-3	1	-1	-3	-2	0	-3	-1	4
5	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
6	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
7	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
8	L	-1	-3	-3	-4	-1	-3	-3	-4	-3	2	2	-3	1	3	-3	-2	-1	-2	0	3
9	L	-1	-3	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	2
10	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
11	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
12	A	5	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
13	W	-2	-3	-4	-4	-2	-2	-3	-4	-3	1	4	-3	2	1	-3	-3	-2	7	0	0
14	A	3	-2	-1	-2	-1	-1	-2	4	-2	-2	-2	-1	-2	-3	-1	1	-1	-3	-3	-1
15	A	2	-1	0	-1	-2	2	0	2	-1	-3	-3	0	-2	-3	-1	3	0	-3	-2	-2
16	A	4	-2	-1	-2	-1	-1	-1	3	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	-1
...																					
37	S	2	-1	0	-1	-1	0	0	0	-1	-2	-3	0	-2	-3	-1	4	1	-3	-2	-2
38	G	0	-3	-1	-2	-3	-2	-2	6	-2	-4	-4	-2	-3	-4	-2	0	-2	-3	-3	-4
39	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-3	-2	0
40	W	-3	-3	-4	-5	-3	-2	-3	-3	-3	-3	-2	-3	-2	1	-4	-3	-3	12	2	-3
41	Y	-2	-2	-2	-3	-3	-2	-2	-3	2	-2	-1	-2	-1	3	-3	-2	-2	2	7	-1
42	A	4	-2	-2	-2	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	1	0	-3	-2	0
...																					

FIGURE 5.10. *Portion of a PSSM from a PSI-BLAST search using RBP4 (NP_006735) as a query. The 199 amino acid residues of the query are represented in rows; the 20 amino acids are in columns. Note that for a given residue such as alanine the score can vary (compare A14, A15, and A16, which receive scores of 3, 2, and 4). The tryptophan in position 40 is invariant in several hundred lipocalins. Compare the score of W40, W3, or W5 (each receives +12) with W13 (+7); in the W3, W5, and W40 positions a match is rewarded more highly, and the penalties for mismatches are substantially greater. A PSSM such as this one allows PSI-BLAST to perform with far greater sensitivity than standard blastp searches.*

columns that required the insertion of a gap into the query sequence were ignored (Altschul et al., 1997), ensuring that the matrix length remains the same as the query length. Some columns of the multiple sequence alignment matrix might have many matches to the query, and others could have none. A unique scoring matrix (profile) is derived from the multiple sequence alignment. Scores are derived for each position in the form $\log(Q_i/P_i)$, where Q_i is the estimated probability for residue i to be found in that column position and P_i is the background probability for that residue.

The unique profile that PSI-BLAST identifies is next used to perform an iterative search. Press the button "run PSI-BLAST iteration 2." The search is repeated using the customized profile, and new proteins are often added to the alignment. This is seen in the second iteration (Fig. 5.11) as the number of database hits rises to 173, including 48 new members above threshold (Table 5.4). Occasionally, protein alignments fall below threshold as the PSSM is modified.

Continue the iteration process until the program message "Converged!" appears. This indicates that no more database matches are found, and the PSI-BLAST search is ended.

What did this search achieve? After a series of position-specific iterations, hundreds of additional database matches were identified. Many distantly related proteins are now shown in the alignment. We can understand how the sensitivity of the search increased by examining the pairwise alignment of the query (RBP4) with a match, mouse apolipoprotein D (Fig. 5.12). In the first PSI-BLAST iteration, the bit score was 46, the expect value was 2e-04 (i.e., 2×10^{-4}), and there were 40 identities and 37 gaps across an alignment of 150 residues. After the second and third iterations, the bit score continued to rise (to 140, then to 159, bits), the E value dropped (to 10^{-32}, then to 10^{-38}), the length of the alignment increased (to 176, then it diminished slightly to 170 residues), and the number of gaps decreased greatly. One notable change occurred in the second iteration, when the amino termini of the two proteins were included in the alignment (Figs. 5.12*a*, *b*). Another notable change occurred in third alignment when the altered profile permitted the closure of several large gaps.

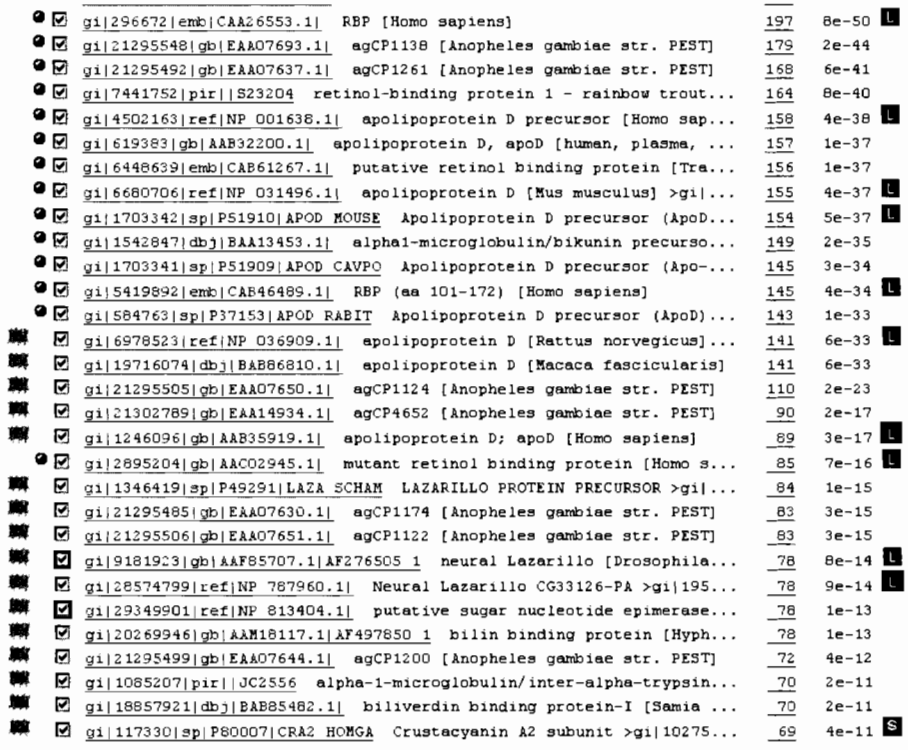


FIGURE 5.11. *Portion of the output of a PSI-BLAST search (iteration 2). Database hits that were present in the previous iteration are shown (ball, left side) while newly added sequences are indicated. Note that these include proteins with the name lipocalin, suggesting that they are authentic RBP homologs.*

We can visualize the PSI-BLAST process by imagining each lipocalin in the database as a point in space (Fig. 5.13). An initial search with RBP4 detects other RBP homologs as well as several apolipoprotein D proteins. The PSSM of PSI-BLAST allows the detection of other lipocalins related to apolipoprotein D. Rat odorant-binding protein is not detected by a RBP blastp search, but it is found by PSI-BLAST.

TABLE 5-4 Results of PSI-BLAST Search with Human Retinol-Binding Protein (NP_006735)

Iteration 1 represents the initial search results, comparable to a standard blastp search result. Subsequent PSI-BLAST iterations provide a dramatic rise in the number of database hits above threshold ($E = 0.005$). Further iterations do not result in substantially different results than seen in iteration 12

Iteration	Number of Hits	Hits Above Threshold	Newly Added	Fell Below Threshold
1	104	49	–	–
2	173	96	48	1
3	236	178	85	3
4	301	240	64	2
5	344	283	43	0
6	342	298	17	2
7	378	310	12	0
8	382	320	11	1
9	391	330	11	1
10	372	336	8	2
11	370	337	2	1
12	371	338	1	0

(a)
```
Score = 46.2 bits (108), Expect = 2e-04
Identities = 40/150 (26%), Positives = 70/150 (46%), Gaps = 37/150 (24%)

Query: 27  VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
           V+ENFD  ++ G WY + +K P      + I A +S+ E G +    K         ++
Sbjct: 33  VQENFDVKKYLGRWYEI-EKIPASFEKGNCIQANYSLMENGNIEVLNK---------ELS 82

Query: 87  ADMVGTF---------TDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCR 137
            D  GT          ++  +PAK +++++ +          +WI+ TDY+ YA+ YSC
Sbjct: 83  PD--GTMNQVKGEAKQSNVSEPAKLEVQFFPLMP-----PAPYWILATDYENYALVYSCT 135

Query: 138 ----LLNLDGTCADSYSFVFSRDPNGLPPE 163
               L ++D      + ++  R+P  LPPE
Sbjct: 136 TFFWLFHVD------FFWILGRNPY-LPPE 158
```

(b)
```
Score =  140 bits (353), Expect = 1e-32
Identities = 45/176 (25%), Positives = 78/176 (43%), Gaps = 33/176 (18%)

Query: 4   VWALLLLAAWAAAERDCRVSSF--------RVKENFDKARFSGTWYAMAKKDPEGLFLQD 55
           V  L+ LA  A      +  +F          V+ENFD  ++ G WY + +K P      +
Sbjct: 2   VTMLMFLATLAGLFTTAKGQNFHLGKCPSPPVQENFDVKKYLGRWYEI-EKIPASFEKGN 60

Query: 56  NIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMV---GTFTDTEDPAKFKMKYWGVASF 112
            I A +S+ E G +    K       + D   + V      ++  +PAK +++++ +
Sbjct: 61  CIQANYSLMENGNIEVLNKEL-----SPDGTMNQVKGEAKQSNVSEPAKLEVQFFPL--- 112

Query: 113 LQKGNDDHWIVDTDYDTYAVQYSCR----LLNLDGTCADSYSFVFSRDPNGLPPEA 164
                +WI+ TDY+ YA+ YSC      L ++D      + ++  R+P  LPPE
Sbjct: 113 --MPPAPYWILATDYENYALVYSCTTFFWLFHVD------FFWILGRNPY-LPPET 159
```

(c)
```
Score =  159 bits (404), Expect = 1e-38
Identities = 41/170 (24%), Positives = 69/170 (40%), Gaps = 19/170 (11%)

Query: 3   WVWALLLLAAWAAAERD--------CRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQ 54
            V  L+ LA  A               +  S  V+ENFD  ++ G WY + K
Sbjct: 1   MVTMLMFLATLAGLFTTAKGQNFHLGKCPSPPVQENFDVKKYLGRWYEIEKIPASFE-KG 59

Query: 55  DNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQ 114
           + I A +S+ E G +    K           V  +      ++  +PAK +++++ +
Sbjct: 60  NCIQANYSLMENGNIEVLNKELSPDGTMNQVKGE--AKQSNVSEPAKLEVQFFPL----- 112

Query: 115 KGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEA 164
                +WI+ TDY+ YA+ YSC             + ++  R+P  LPPE
Sbjct: 113 MPPAPYWILATDYENYALVYSCTTFFWL--FHVDFFWILGRNPY-LPPET 159
```

FIGURE 5.12. *PSI-BLAST search detects distantly related proteins using progressive iterations with a PSSM. (a) A search with RBP4 as a query (NP_006735) detects the lipocalin mouse apolipoprotein D (NP_031496) in the first iteration. As the search progresses to (b) the second iteration and (c) the third iteration, the bit score becomes higher, the expect value decreases, and the number of gaps in the alignment decreases.*

As another example of the usefulness of PSI-BLAST, consider human vacuolar protein-sorting protein 45 (*h-vps45*; NP_009189), a member of the Sec1 trafficking family. A blastp nr search results in 66 hits having an *E* value below 0.002. Additional hits with a value worse than 0.002 do not appear to be relevant homologs of this protein. Perform a PSI-BLAST search. Initially, there are 69 hits (comparable to the blastp nr result). But the first iteration yields 87 hits, the second iteration yields 94 hits, and convergence is reached on the third iteration (with 91 hits). By inspection, these are authentically related proteins.

The number of iterations that a PSI-BLAST search performs relates to the number of hits (sequences) in the database that running the program reports. After each PSI-BLAST iteration, the results that are returned describe which sequences match the input PSSM.

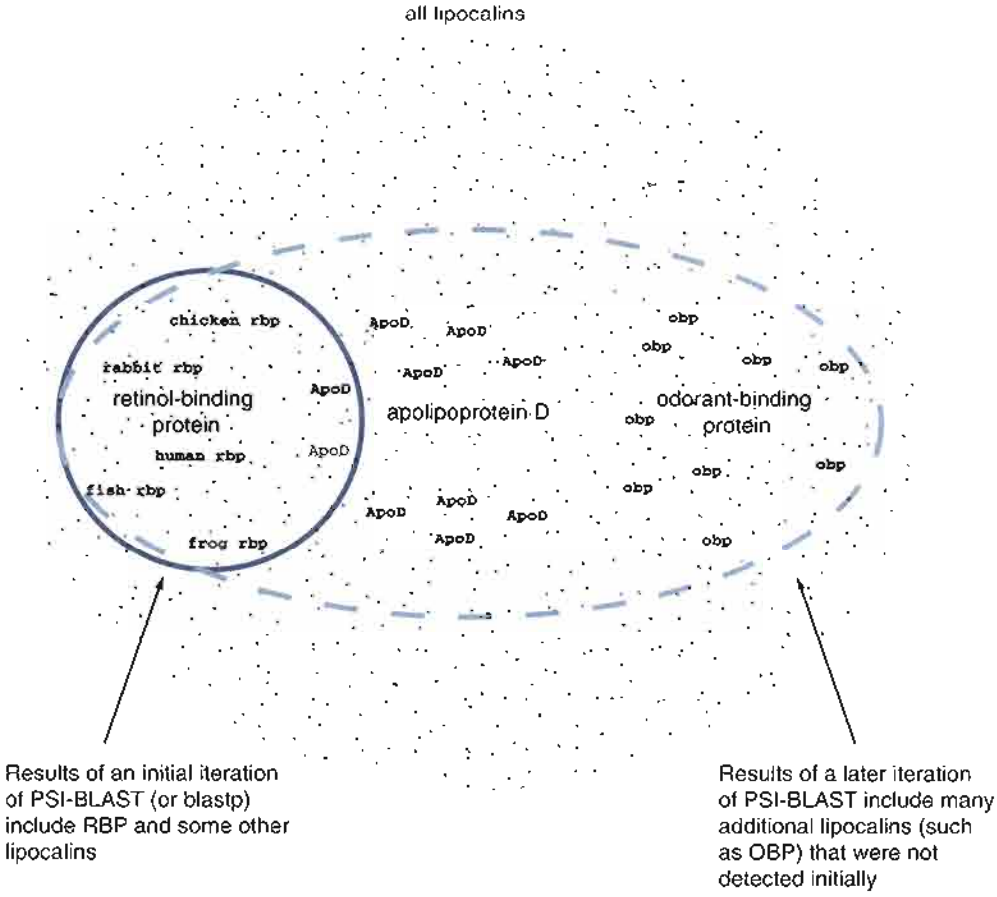


FIGURE 5.13. PSI-BLAST algorithm increases the sensitivity of a database search by detecting homologous matches with relatively low sequence identity. In this figure, each dot represents a single protein, some of which are labeled RBP (retinol-binding protein), ApoD (apolipoprotein D), or OBP (odorant-binding protein). All these proteins are homologous by virtue of their membership in the lipocalin family. A standard blastp search with RBP returns matches that are relatively close to RBP in sequence identity, and the result (represented by the circle at left) may include additional matches to lipocalins such as ApoD. However, many other lipocalins such as OBP are not detected. The fundamental limitation in standard BLAST search sensitivity is the reliance on standard PAM and BLOSUM scoring matrices. In a PSI-BLAST search, a PSSM generates a scoring system that is specific to the group of matches detected using the initial query sequence (e.g., RBP). While the initial iteration of a PSI-BLAST search results in an identical number of database matches as a standard BLAST search, subsequent PSI-BLAST iterations (represented by the dashed oval) using a customized matrix extend the results to allow the detection of more distantly related homologs.

Assessing Performance of PSI-BLAST

There are several ways to assess the performance of PSI-BLAST. When a query is searched against a large database such as SwissProt, the PSSMs can be searched against versions of the database that either are shuffled or have the order of each sequence reversed. When this is done, the PSI-BLAST expect values are not significant (Altschul et al., 1997).

In another approach, several groups have compared the relationships detected using PSI-BLAST to those detected by the rigorous structural analysis of homologous proteins that share limited amino acid identity. Park and colleagues (1998) used the structural classification of proteins (SCOP) database. They found that PSI-BLAST searches of this database were more accurate using an E value of 0.0005 (the

In a related approach, Schaffer et al. (2001) plotted the number of PSI-BLAST false positives versus true positives to generate a "sensitivity curve." They used this plot to assess the accuracy of PSI-BLAST using a variety of adjustments to the parameters.

SCOP (Chapter 9) is available at ▸http://scop.mrc-lmb.cam.ac.uk/scop/. It was developed by Cyrus Chothia and colleagues. Park et al. (1998) used the PDBD40-J database, which contains proteins of known structure with ≤40% amino acid identity.

default inclusion threshold for E is 0.005). They estimated the false-positive rate for PSI-BLAST matches by assessing how many false predictions were made out of 432,680 possible matches. At a low rate of false positives (1 in 100,000), PSI-BLAST detected 27% of homologous matches in the database; at a higher rate of false positives (1 in 1000), PSI-BLAST detected 44% of the homologous matches. This performance is comparable to that observed for SAM-T98, an implementation of a hidden Markov model procedure (Chapter 10), and PSI-BLAST was far more sensitive than the standard gapped BLAST or FASTA algorithms.

Friedberg et al. (2000) used the fold classification based on structure–structure alignment of proteins (FSSP) and Distant Aligned Protein Sequences (DAPS) (see Chapter 9). Their study included several lipocalins: bovine RBP identified bovine odorant-binding protein; bovine RBP detected both mouse major urinary protein and a bilin-binding protein.

Friedberg and colleagues (2000) assessed the accuracy of PSI-BLAST alignments. They selected proteins from two structure databases. Fifty-two sequences out of 123 successfully detected their known structural matches using PSI-BLAST, even though the aligned pairs shared less than 25% amino acid identity and none of the pairs could be detected by the Smith–Waterman algorithm. They then compared the alignments generated by PSI-BLAST with the alignments measured using X-ray crystallography and found that, on average, about 44% of the residues were correctly aligned.

PSI-BLAST Errors: Problem of Corruption

PSI-BLAST is useful to detect weak but biologically meaningful relationships between proteins. The main source of error in PSI-BLAST searches is the spurious amplification of sequences that are unrelated to the query. This problem most often arises when the query (or the profile generated after PSI-BLAST iterations) contains regions with highly biased amino acid composition. Once the program finds even one new protein hit having an E value even slightly above the inclusion threshold, that new hit will be incorporated into the next profile and will reappear in the next PSI-BLAST iteration. If the hit is to a protein that is not homologous to the original query sequence, then the PSSM has been corrupted.

We can define corruption as occurring when, after five iterations of PSI-BLAST, the PSSM produces at least one false-positive alignment of $E < 10^{-4}$. This definition is adapted from Schaffer et al. (2001).

There are three main approaches to stopping corruption of PSI-BLAST queries. (1) You can apply a filtering algorithm that removes biased amino acid regions. These "low-entropy" regions include stretches of amino acids that are highly basic, acidic, or rich in a residue such as proline. The NCBI PSI-BLAST site employs the SEG program for this purpose, applying the filtering algorithm to database sequences that are detected by the query. (2) You can adjust the expect level from its default value (e.g., $E = 0.005$) to a lower value (e.g., $E = 0.0001$). This may suppress the appearance of false positives, although it could also interfere with the detection of true positives. (3) You can visually inspect each PSI-BLAST iteration. As shown in Figure 5.11, you can click on the checkboxes adjacent to each protein listed in the PSI-BLAST output and remove suspicious ones. As an example, your query protein may have a generic coiled-coil domain, and this may cause other proteins sharing this motif (such as myosin) to score better than the inclusion threshold even though they are not homologous.

SEG was described by Wootton and Federhen (1996) and is available at ▸ftp://ncbi.nlm.nih.gov/pub/seg/.

If a protein has several motifs, such as both a kinase domain and a C2 domain, PSI-BLAST may find database matches related to both. The results must be interpreted carefully. One should not conclude that the kinase domain is related to the C2 domain. An example of this type is presented in problem 5.1. For PSI-BLAST searches with multidomain proteins, it may be helpful to search using just one region of interest, such as the reverse transcriptase domain of HIV-1 *pol* protein.

PATTERN-HIT INITIATED BLAST (PHI-BLAST)

Often a protein you are interested in contains a pattern, or "signature," of amino acid residues that help define that protein as part of a family. For example, a signature might be an active site of an enzyme, a string of amino acid residues that define a structural or functional domain of a protein family, or even a characteristic signature of unknown function (such as GXW in the lipocalins). Pattern-hit initiated BLAST (PHI-BLAST; ϕ-BLAST) is another specialized BLAST program that allows you to search with a query and to find database matches that both match a pattern and are significantly related to the query (Zhang et al., 1998). PHI-BLAST may be preferable to simply searching a database with a short query corresponding to a pattern, because such a search could result in the detection of many random matches or proteins that are unrelated to your query protein.

The PHI-BLAST pattern limits only the original database search (and not the PSI-BLAST PSSM search).

PHI-BLAST is launched from the same NCBI web page as PSI-BLAST. These are independent programs, although the output of a PHI-BLAST search can be used as input into PSI-BLAST. PHI-BLAST is performed once (like blastp) and not iteratively (like PSI-BLAST).

Consider a PSI-BLAST search of bacterial sequences using human *RBP4* as a query. The result (as of August 2002) is that there are seven database matches in the first iteration, all of which are bacterial sequences with scores worse than the default threshold of $E = 0.005$. We know that there are bacterial lipocalins distantly related to human *RBP4*; one way we can identify them is to perform an Entrez protein query with the words "bacteria lipocalin." Select two of these bacterial lipocalin protein sequences and align them with RBP4 (Fig. 5.14; we will describe how to make multiple sequence alignments in Chapter 10). This alignment shows us which amino acid residues are actually shared between *RBP4* and the two bacterial proteins. Focusing on the GXW motif that is shared between almost all lipocalins, we can try to define a pattern (or signature) of amino acids that is shared by *RBP4,* the two bacterial lipocalins, and possibly many other bacterial lipocalins. The purpose of defining a signature is to customize the PSI-BLAST algorithm to search for proteins containing that signature.

How is the signature or pattern defined? We do not expect the signature to be exactly identical between all bacterial lipocalins, and so we want to include freedom for ambiguity. We can define any pattern we want; as an example we will examine

```
        1                                                    50
ecb1c   MRLLPLVAAA TAAFLVVACS SPTPPRGVTV VNNFDAKRYL GTWYEIARFD
  vc    MRAIFLILCS V...LLNGCL G..MPESVKP VSDFELNNYL GKWYEVARLD
hsrbp   ~~~MKWVWAL LLLAAWAAAE RDCRVSSFRV KENFDKARFS GTWYAMAKKD
```

FIGURE 5.14. *Choosing a pattern for a PHI-BLAST search. Human RBP4 (accession NP_006735) was multiply aligned with two bacterial lipocalins that were selected using an Entrez query (these are* E. coli *P39281 and* Vibrio cholerae *Q08790). The purpose of evaluating these three protein sequences together is to try to identify a short, sequential pattern of amino acid residues that consistently occurs in a protein family. This pattern then is included in the search to increase its sensitivity and specificity. The alignment was performed using the GCG PileUp program, and a portion of the alignment is shown above. By inspection, the invariant GXW motif that is typical of lipocalins is evident (shaded red). A PHI pattern can be selected that includes these residues and several more. As an example, we select the pattern GXW [YF] [EA] [IVLM], in which, following GXW, the next position contains either Y or its closely related residue F, then an E or A because both are present in these proteins, and finally the hydrophobic residues I, V, M, and L. The user can select any pattern by trial and error. The database will then be searched, with a requirement that all database matches include the selected pattern.*

The syntax for a PHI-BLAST pattern is derived from the Prosite dictionary (Chapter 8) and is described at ▸http://www.ncbi.nlm.nih.gov/blast/html/PHIsyntax.html.

the multiple alignment in Fig. 5.14 and create the pattern GXW[YF] [EA] [IVLM]. Note that the pattern you choose must not occur too commonly; the algorithm only allows patterns that are expected to occur less frequently than one time for every 5000 database residues. In general, it is acceptable to choose any pattern with four completely specified residues or three residues with average background frequencies of ≤5.8% (Zhang et al., 1998).

The BLAST search output is restricted to a subset of the database consisting of proteins that contain that specified pattern. By inspection of the pattern we have chosen, each database match must have a G; the X allows any residue to come next; the W specifies that the third amino acid residue of the pattern must be a W. Next, we write [YF] to specify that the next amino acid must be either a Y or an F; we choose this because it is very common for tyrosines to be substituted with phenylalanine (see Chapter 3). In the next position, we select [EA] because these two residues are used in the alignment of Figure 5.14. Finally, we select [IVLM] to correspond to the three residues in the alignment (I, V, M) as well as an additional hydrophobic residue (L) that we add based on the intuition that any hydrophobic residue might occur in this position.

We next use PHI-BLAST by going to the PSI-BLAST page (from ▸http://www.ncbi.nlm.nih.gov/BLAST). Under query options notice the box called "PHI pattern." Enter the pattern GXW[YF][EA][IVLM] (Fig. 5.15). The result of this search is that instead of 5 matches (none better than threshold) there are now dozens of database matches, including six bacterial lipocalins having scores better than threshold. The ensuing PSI-BLAST iteration, which no longer uses the PHI pattern but instead uses a search-specific PSSM, will successfully identify a large family of bacterial lipocalins.

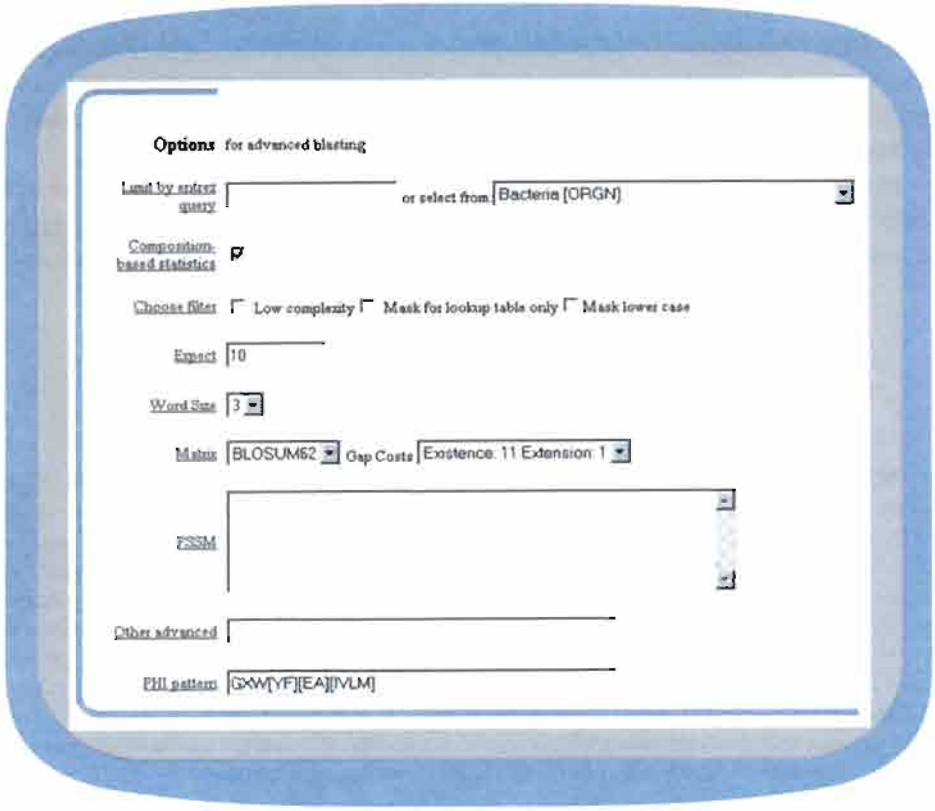


FIGURE 5.15. PHI-BLAST search uses a pattern of several amino acids as a signature to guide the database search. The syntax of the PHI pattern is the same as that used by PROSITE (Chapter 8). After the user inputs this pattern into the options, the equivalent of a PSI-BLAST search is performed. The addition of a PHI pattern often results in an increased number of database matches. Choosing an appropriate pattern to input into a PHI-BLAST search can be performed by trial and error or by selecting a region that appears to be well conserved in a protein family.

```
>gi|2497702|sp|Q46036|BLC_CITFR   OUTER MEMBRANE LIPOPROTEIN BLC PRECURSOR
 gi|2121019|pir||I40710   outer membrane lipoprotein - Citrobacter freundii
 gi|717136|gb|AAC46456.1|   lipocalin precursor [Citrobacter freundii]
          Length = 177

 Score = 17.3 bits (52), Expect = 0.002
 Identities = 23/82 (28%), Positives = 36/82 (43%), Gaps = 1/82 (1%)

Query:  27  VKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVC 86
pattern 38              ******
            V  NFD  R+ GTWY +A+ D       D + A +S+ + G ++   KG       W
Sbjct:  30  VVNNFDAKRYLGTWYEIARFDHRFERGLDKVTATYSLRDDGGINVINKGYNPDREMWQK- 88

Query:  87  ADMVGTFTDTEDPAKFKMKYWG 108
             +    FT     A  K+ ++G
Sbjct:  89  TEGKAYFTGDPSTAALKVSFFG 110
```

FIGURE 5.16. *Results of a PHI-BLAST search resemble a standard BLAST or a PSI-BLAST search, but additional information about the selected pattern are included. Here, the region of both the query (human* RBP4, *NP_006735) and a database match (subject; a bacterial lipocalin, Q46036) are aligned to the PHI pattern that was selected (GXW [YF] [EA] [IVLM]).*

The output of the PHI-BLAST search is identical to the PSI-BLAST output, except that information about where both the query and each database sequence match the PHI pattern is shown (Fig. 5.16, asterisks). The PHI-BLAST algorithm employs a statistical analysis based on identifying alignment A_0 spanned by the input pattern and regions A_1 and A_2 to either side of the pattern, which are scored by gapped extensions. Scores S_0, S_1, and S_2 corresponding to these regions are calculated, and PHI-BLAST scores are ranked by the score $S' = S_1 + S_2$ (ignoring S_0). The alignment statistics are closely related to those used for blastp searches (Zhang et al., 1998).

USING BLAST FOR GENE DISCOVERY

A common problem in biology is finding a new gene. Traditionally, genes and proteins were identified using the techniques of molecular biology and biochemistry. Complementary DNAs were cloned from libraries, or proteins were purified then sequenced based upon some biochemical criteria such as enzymatic activity. Such experimental biology approaches will always remain essential. Bioinformatics approaches can also be useful to provide evidence for the existence of new genes. For our purposes a "new" gene refers to the discovery of some DNA sequence in a database that is not annotated (described). You may want to find new genes for many reasons:

- You want to study a lipocalin that no one has characterized before, perhaps in a specific organism of interest such as a plant or archaeon.
- You are interested in the lipocalins, and you see that one has been described in the tears of hamsters. Could there be a new, undiscovered gene that encodes a lipocalin protein expressed in human tears? (At present, there is one!)
- You want to know if bacteria have lipocalins. If so, this might give you insight into the evolution of this family of carrier proteins.
- You study diseases in which sugars are not processed properly, and as part of this research, you study sugar transport in human cell lines. You know that glucose transporters have been characterized by biochemical assays (e.g., sugar uptake). You also know that there is a family of glucose transporter genes (and proteins) that have been deposited in GenBank. You cloned all the known transporters, expressed them in cells, and found that none of the recombinant proteins transports your sugar. You hypothesize that there must be at least one more transporter that has not yet been described. Is there a way to search the database to find human genes encoding novel transporters?

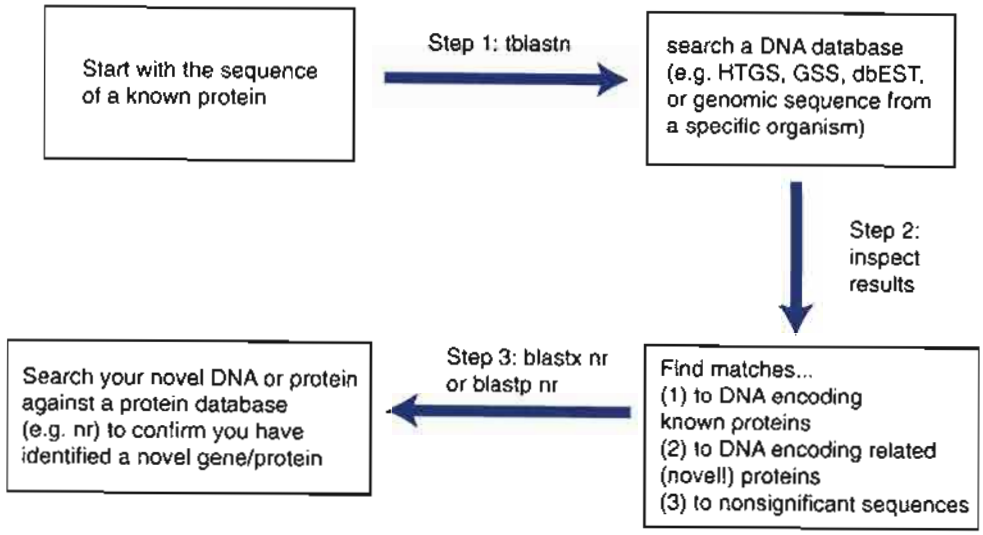


FIGURE 5.17. How to discover a novel gene by BLAST searching. Begin with the sequence of a known protein such as RBP. Perform a tblastn search of a DNA database. It is unlikely that there are many "novel" genes in the well-characterized genomes of organisms such as human, yeast, or E. coli. *Thus it may be helpful to search databases of organisms that are poorly characterized or not fully annotated. The tblastn search may result in two types of significant matches: (1) matches of your query to known proteins that are already annotated and (2) homologous proteins that have not yet been annotated ("novel" genes and corresponding novel proteins). (3) The DNA sequence corresponding to the putative novel gene may be searched using the blastx algorithm against the nonredundant (nr) database. This may confirm that the DNA does indeed encode a protein that has no perfect match to any described protein.*

- You are studying the HIV pol protein, in particular its reverse transcriptase domain. You would like to identify an example of this domain in a eukaryotic protein. However, rather than studying a known eukaryotic protein with this motif, you would like to study a novel one that has never been characterized.

A general strategy to solve any of these problems is presented in Figure 5.17. With practice, this entire process can be completed in just a matter of minutes. The starting point is a known protein sequence that is used as a query in a database search (Fig. 5.17, step 1). The best database to search is one that contains ESTs, genomic sequence, or other DNA sequence. It is usually less effective to search a database such as yeast (*S. cerevisiae*), *E. coli,* or any of the nonredundant DNA or protein databases, because in all these cases the sequences are relatively well annotated, and it is difficult to identify "novel" genes. One useful strategy is to perform a tblastn search using your protein query against a DNA database such as those listed in Table 5.1.

The next procedure is to carefully inspect your tblastn results (Fig. 5.17, step 2). With experience, it becomes possible to identify three kinds of results:

1. Some database hits match the query exactly or nearly exactly. If the query is human *RBP4* and the database is not human, the match may be an RBP4 ortholog that has already been characterized. This would not represent a novel finding.
2. Some database matches match the query significantly, but these proteins encoded by the DNA in the database are not annotated as known proteins. Sequences in this category are candidate novel genes.
3. Some database matches are not significant. It requires experience to learn which database matches are authentic and which are not. Authentic

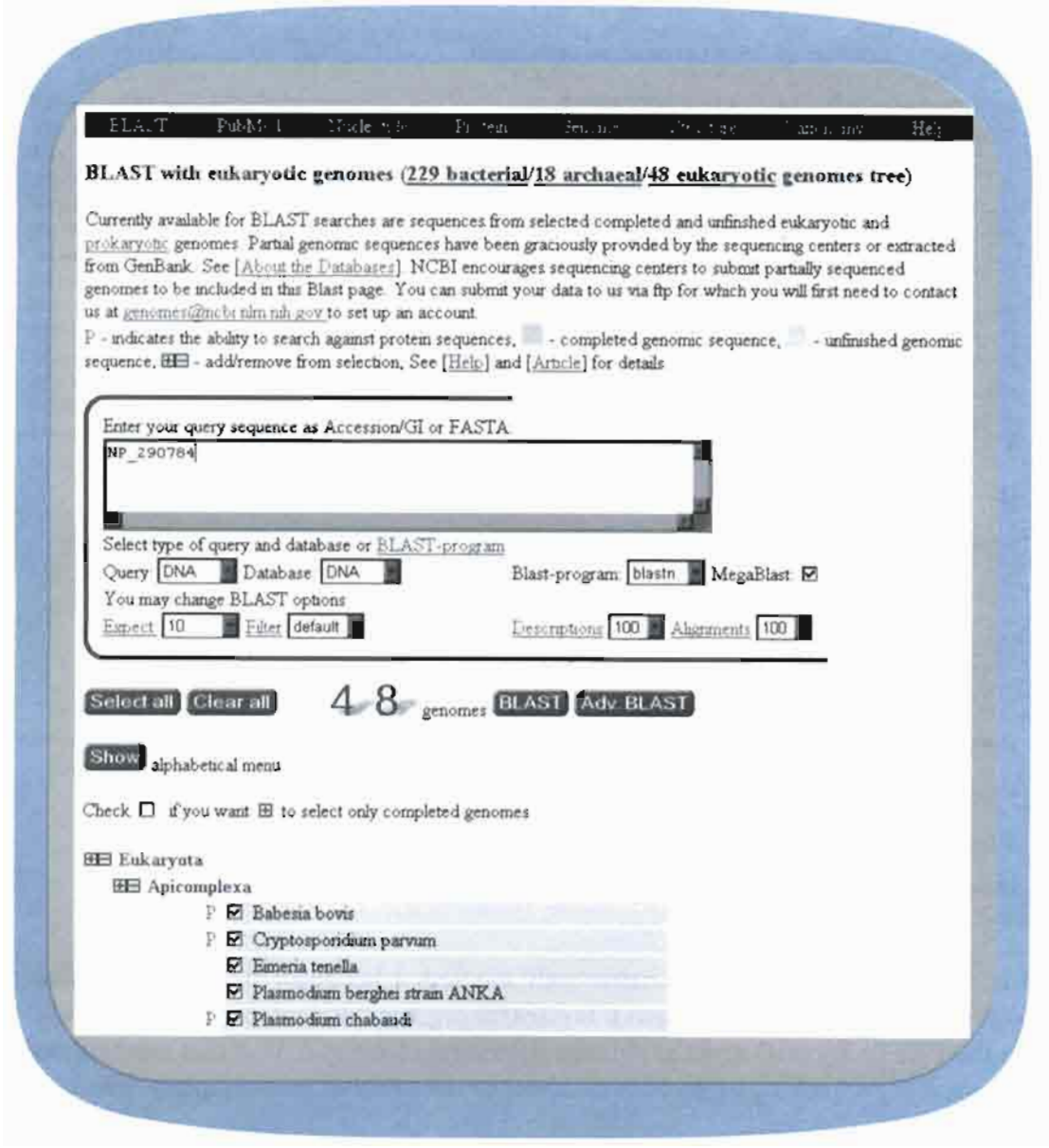


FIGURE 5.18. Procedure to find a novel gene by BLAST searching. In this case, use the "other eukaryotes" genomic BLAST site (linked from ▸http://www.ncbi.nlm.nih.gov/BLAST), enter the human RBP4 accession number, set the program to tblastn, and select the organism of choice. Clicking "select all" provides access to the genomes of several dozen organisms, such as Aspergillus fumigatus, Entamoeba histolytica, Trypanosoma brucei, *and* Plasmodium falciparum. *The purpose of this search is to identify proteins related to RBP4 that are predicted to be encoded by uncharacterized genes in unfinished genomes.*

homologs are likely to be approximately the same size (in amino acids) as the query, and they are likely to contain the same motifs, such as the GXW signature that is typical of lipocalins.

To find a novel lipocalin, enter an *E. coli* lipocalin accession (NP_290784) into a search of unfinished eukaryotic genomes (at ▸http://www.ncbi.nlm.nih.gov/BLAST). Set the database to all; there are many dozens of eukaryotic genomes represented (Fig. 5.18). (Alternatively, one could select other databases to search, such as those in Table 5.1.) The result of the search is that a group of predicted proteins have significant matches to the bacterial query sequence (Fig. 5.19). One of these is a *Dictyostelium discoideum* (slime mold) protein. This may be a novel protein in the sense that no one has identified this gene (or the corresponding protein) in the *Dictyostelium* genome. We can assess whether it is indeed novel by performing a database search (Fig. 5.17, step 3). The result of a blastp search using the *Dictyostelium* protein shows that it is closely related to several other bacterial lipocalins, but it has not been annotated (i.e., explicitly described) in the GenBank database (Fig. 5.20).

A similar strategy to that described in Figure 5.15 can be used to find genes from any organism. (However, as some genomes such as the human, *S. cerevisiae,*

(*a*)

```
                                                                    Score     E
Sequences producing significant alignments:                         (bits)  Value

gnl|ti 7165|44877221  19866907818432                                 176    5e-43
gnl|DictyConsortium 44689|jena JC1c122d04.s1  Clone JC1c122d...      149    4e-35
gnl|DictyConsortium 44689|jena JC1c276c01.r1  Clone JC1c276c...      116    4e-25
gnl|DictyConsortium 44689|jena JC1a241a10.s2  Clone JC1a241a...       99    7e-20
gnl|DictyConsortium 44689|jena JC1c236h06.r1  Clone JC1c236h...       94    3e-18
gnl|DictyConsortium 44689|iibsp.fa IIBCP1D0380  Dictyosteliu...      82    1e-14
gnl|ti 7165|47325132  19866908692623                                 68    2e-10
gnl|ti 7165|57669632  19866913818355                                 68    2e-10
gnl|ti 7165|55759626  19866909741201                                 63    5e-09
gnl|ti 7165|55559725  19866910183296                                 63    7e-09
gnl|DictyConsortium 44689|jena JC1b180c09.s1  Clone JC1b180c...      63    7e-09
gnl|ti 7165|60399996  19866914851608                                 61    3e-08
gnl|DictyConsortium 44689|jena JC1c45b10.r1  Clone JC1c45b10...      60    3e-08
gnl|DictyConsortium 44689|jena JC1b154f03.r1  Clone JC1b154f...      59    1e-07
```

(*b*)

```
>gnl|DictyConsortium 44689|jena JC1c276c01.r1 Clone JC1c276c01, reverse read,
          2001-01-30
          Length = 474

 Score =  116 bits (291), Expect = 4e-25
 Identities = 58/116 (50%), Positives = 75/116 (64%)
 Frame = +2

Query: 44  YEIARFDHRFERGLEKVTATYSLRDDGGLNVINKGYNPDRGMWQQSEGKAYFTGAPTRAA 103
           YEIAR +  +ER +  V+A Y+L  DG ++VIN GYN      +   GKA F      A
Sbjct: 116 YEIARLNFYYERDMNNVSAEYTLNKDGTISVINSGYNYVEKKRESLNGKALFVHGSHEAM 295

Query: 104 LKVSFFGPFYGGYNVIALDREYRHALVCGPDRDYLWILSRTPTISDEVKQEMLAVA 159
           LKVSFFGPFY GYNVIA+D +Y++ALV G    YLWILSR  +I D +K++ L +A
Sbjct: 296 LKVSFFGPFYAGYNVIAIDPDYKYALVAGRSLHYLWILSRETSIPDHIKKQYLQLA 463
```

FIGURE 5.19. *(a) Result of a tblastn search of unfinished genomes. (b) An alignment is shown. This D. discoideum protein is predicted to be encoded from a genomic DNA fragment. Copy the slime mold protein sequence; to determine if this is a "novel" protein, we will use it as a query in a blastp search of the nr database.*

and other model genomes are intensively studied, it becomes increasingly difficult to discover novel genes in them.) Try searching the TIGR databases using RBP4 (in the FASTA format) and the tblastn algorithm (see Fig. 5.3). If this method does not work, try searching a different database or use another protein query of interest.

PERSPECTIVE

While BLAST searching has emerged as a fundamental tool for studying proteins and genes (Chapter 4), many specialized BLAST applications have also been developed. These applications include variant algorithms (such as the PSSM of PSI-BLAST) and specialized databases (such as a variety of organism-specific databases). PSI-BLAST has been used extensively to characterize proteins encoded by complete genomes (Chapters 14–17). PSI-BLAST can only be successfully applied to cases in which a blastp search results in at least some statistically significant result.

```
☐>gi|23137293|ref|ZP_00119001.1|   hypothetical protein [Cytophaga hutchinsonii]
          Length = 179

 Score =  127 bits (319), Expect = 3e-29
 Identities = 66/116 (56%), Positives = 86/116 (74%)

Query: 1   YEIARLNFYYERDMNNVSAEYTLNKDGTISVINSGYNYVEKKRESLNGKALFVHGSHEAM 60
           YEIARL++ +E+++NNV+A Y+L +DG I V N GYN  ++K E   GKA  V    E
Sbjct: 46  YEIARLDYSWEKNLNNVTATYSLREDGKIKVDNKGYNVKKEKWEESVGKAAPVADPAEGR 105

Query: 61  LKVSFFGPFYAGYNVIAIDPDYKYALVAGRSLHYLWILSRETSIPDHIKKQYLQLA 116
           LKVSFFGPFYAGYNV+AID  Y YALVAG +  YLWILSR+ +IP+++K+ YL+ A
Sbjct: 106 LKVSFFGPFYAGYNVVAIDDAYTYALVAGENTKYLWILSRKKTIPENVKEAYLKKA 161
```

FIGURE 5.20. *The* Dictyostelium *lipocalin is "novel" based upon a blastp search. The best database matches are to lipocalins from other organisms (such as Cytophaga hutchinsonii). Thus, this search strategy has resulted in the successful identification of a lipocalin in an organism that has not been annotated and deposited in the main (nr) GenBank database.*

The exponential rise in DNA sequence data (Fig. 2.1) presents us with massive amounts of information about genes and proteins. BLAST searching is a fundamental tool for searching these databases. A BLAST search is often more definitive than a literature search for answering questions about protein or gene families across the tree of life. In this chapter, we described several ways to use alternative BLAST databases and alternative BLAST algorithms to perform database searches. These tools will continue to be fundamentally important to biology for many years to come, especially as the pace of genomic sequencing continues to accelerate.

PITFALLS

As with any bioinformatics problem, it is essential to define the purpose of a database search. What are you trying to accomplish? Once you have decided this, you can select the appropriate database and search algorithm.

For PSI-BLAST, the biggest problem is obtaining false positives. Once a spurious sequence has been detected that is better than some expect value cutoff, it will be included in the PSSM for the next iteration. This iteration will almost certainly find the spurious sequence again and will probably expand the number of database matches. To avoid this problem:

- Inspect the results for apparently spurious database matches. If you see them, remove such spurious matches by deselecting them.
- Adjust the expect value as appropriate.
- Perform "reverse" searches in which you evaluate a potentially spurious PSI-BLAST result by using that sequence as a query in a BLAST search.
- Further evaluate a marginal database match by performing pairwise sequence alignment as described in Chapter 3.

For PHI-BLAST, the most common problem encountered is that new users do not have a feel for the rules involved in creating a PHI-BLAST pattern. The best approach is to practice using a variety of signatures.

WEB RESOURCES

We have described three kinds of BLAST and related search tools:

- Organism-specific databases for BLAST searching (Table 5.1)
- BLAST sites that focus on specialized molecules (Table 5.2)
- Alternative algorithms for database searching (Table 5.3) including PSI-BLAST (Table 5.5)

TABLE 5-5 Servers Offering PSI-BLAST

Site	URL
SBASE 9.0 (Trieste)	▸http://www3.icgeb.trieste.it/~sbasesrv/blast41n.html
NIH (Japan)	▸http://spiral.genes.nig.ac.jp/homology/psi_blaste.shtml
Pasteur Institute	▸http://bioweb.pasteur.fr/seqanal/interfaces/psiblast.html
CMBI, Netherlands	▸http://www.cmbi.kun.nl/bioinf/tools/psiblast.shtml
Human Genome Center, University of Tokyo	▸http://blue1.ims.u-tokyo.ac.jp:8080/psiblast.html

Discussion Questions

[5-1] BLAT is an extremely fast, accurate program. Why will it not replace BLAST or at least become as commonly used as BLAST?

[5-2] In the original implementation of PSI-BLAST, the algorithm performed a multiple sequence alignment and deleted all but one copy of aligned sequence segments having ≥98% identity (Altschul et al., 1997). In a recent modification, the program now purges segments having ≥94% identity. What do you think would happen if this percentage were adjusted to ≥75%? How could you test this idea in practice?

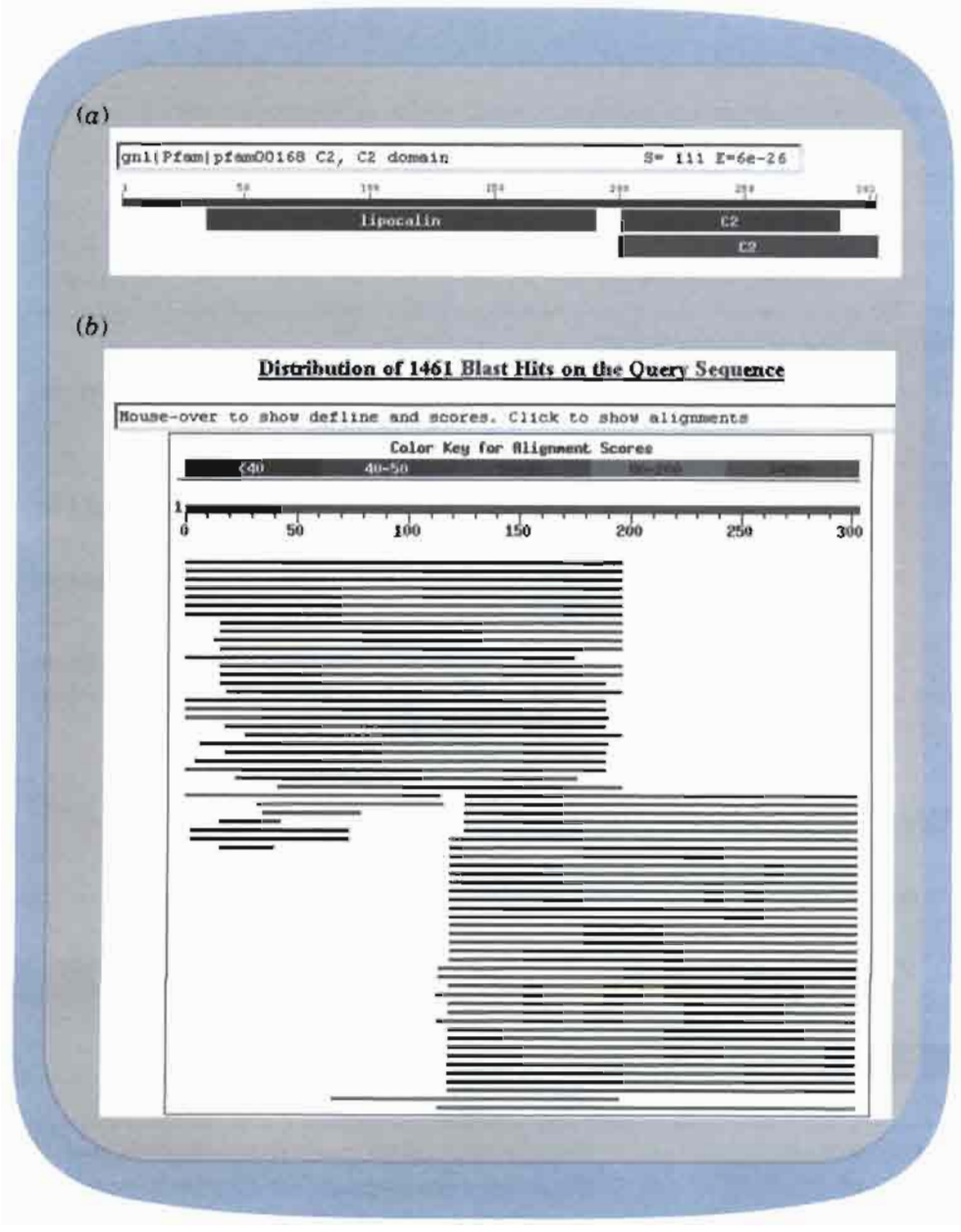


FIGURE 5.21. *(a) After creating a hybrid protein consisting of human* RBP4 *and the C2 domain of human protein kinase Cα, the PSI-BLAST formatting page shows that both domains have been detected. (b) The result of the third PSI-BLAST iteration shows that this algorithm effectively returns database matches corresponding to both domains. It would be incorrect to presume that any lipocalin is homologous to any C2 domain-containing protein. In general, all that PSI-BLAST results provide is a list of matches to a PSSM.*

PROBLEMS

[5-1] Create an artificial protein sequence consisting of human RBP4 followed by the C2 domain of human protein kinase Cα. Enter this combined sequence into a PSI-BLAST search. The result is shown in Figure 5.21. In general, are multiple domains always detected by the PSI-BLAST program? Do any naturally occurring proteins have both lipocalin and C2 domains?

[5-2] The malarium parasite *Plasmodium vivax* has a multigene family called *vir* that is specific to that organism (del Portillo et al., 2001). There are 600–1000 copies of these genes, and they may have a role in causing chronic infection through antigenic variation. Select *vir1* and perform a blastp search of the nonredundant database. Then perform a PSI-BLAST search with the same entry.

(a) In an initial search, approximately how many proteins have an *E* value less than 0.002, and how many have a score greater than 0.002?

(b) What is the score of the best new sequence that is added between the first iteration and the second iteration of PSI-BLAST?

[5-3] We previously performed a series of BLAST searches using HIV-1 pol as a query (NP_057849). Perform a blastp search using this query. Look at the taxonomy report to see which viruses match this query. Next, repeat the search using several iterations of PSI-BLAST. Compare this taxonomy report to that of the blastp search. What do you observe? Are there any nonviral sequences detected in the PSI-BLAST search? Did you expect to find any?

[5-4] Huntington disease is a neurodegenerative disorder caused by a triplet repeat expansion of the nucleotides CAG in the huntingtin gene. This causes an expansion of a string of glutamine residues in the huntingtin gene. The function of huntingtin is poorly understood. Orthologs have been described in several eukaryotic species, but not in bacteria. Identifying a bacterial homolog could be important in order to learn both the function of huntingin (the two proteins could share a similar function) and the evolution of the proteins.

First, perform a standard blastp search of the nonredundant (nr) database using human huntingtin protein as a query. This will tell you whether any obvious bacterial protein homologs are present in the database. Then use PSI-BLAST and/or PHI-BLAST to decide if there are any bacterial homologs. You may select a "signature" and perform a PHI-BLAST search to try to extend the number of database matches. Note that this is a difficult problem for which no solution has been described in the literature, and there are many possible approaches.

SELF-TEST QUIZ

[5-1] Raw DNA sequences (other than Ref seq) in the EMBL and NCBI databases:

(a) Overlap entirely

(b) Overlap to a substantial degree but have distinct sequences

(c) Have relatively little overlap

[5-2] A PSI-BLAST search is most useful when you want to do the following:

(a) Find the rat ortholog of a human protein.

(b) Extend a database search to find additional proteins.

(c) Extend a database search to find additional DNA sequences.

(d) Use a pattern or signature to extend a protein search.

[5-3] Which of the following BLAST programs uses a signature of amino acids to find proteins within a family?

(a) PSI-BLAST

(b) PHI-BLAST

(c) MS BLAST

(d) WormBLAST

[5-4] Which of the following BLAST programs is best used for the analysis of immunoglobulins?

(a) RPS-BLAST

(b) PHI-BLAST

(c) IgBLAST

(d) ProDom

[5-5] In a position-specific scoring matrix, the column headings can have the 20 amino acids, and the rows can represent the residues of a query sequence. Within the matrix, the score for any given amino acid residue is assigned based on:

(a) A PAM or BLOSUM matrix

(b) Its frequency of occurrence in a multiple sequence alignment

(c) Its background frequency of occurrence

(d) The score of its neighboring amino acids

[5-6] As part of a PSI-BLAST search, a score is assigned to an alignment between a query sequence and a database match over some length (such as 50 amino acid residues). It is possible for this pairwise alignment to receive a higher or lower score over successive PSI-BLAST iterations, even though there is no change in which amino acid residues are aligned.

(a) True

(b) False

[5-7] A position-specific scoring matrix is said to be "corrupted" when it incorporates a spurious sequence (i.e., a false-positive result). Which of the following choices is the best way to reduce corruption?

(a) Lower the *E* value.

(b) Remove filtering.

(c) Use a shorter query.
(d) Run fewer iterations.

[5-8] What is the relationship between PSI-BLAST and PHI-BLAST?
(a) They are launched from the same web page at NCBI but are otherwise unrelated kinds of BLAST searches.
(b) A PHI-BLAST search initiates a PSI-BLAST search but constrains the database matches to those containing a user-specified, short amino acid pattern.
(c) They both employ a position-specific scoring matrix.
(d) They both involve successive search iterations.

[5-9] If you want to find proteins that are distantly related to your query, which of these strategies is most likely to be successful?
(a) Using PHI-BLAST, because you can specify a signature that is selective for the proteins related to your query
(b) Using PSI-BLAST, because its strategy of using a position-specific scoring matrix is likely to be most sensitive
(c) Using BLASTP, because you can adjust the scoring matrices to maximize the sensitivity of your search
(d) Using organism-specific databases, because they are most likely to include distantly related sequences

[5-10] Which of the following steps is crucial to validating a sequence you believe to be that of a novel gene?
(a) Performing a PSI-BLAST search
(b) Checking the EST database to see where this gene might be expressed
(c) Checking LocusLink to see if other family members of this gene have been annotated
(d) BLAST searching your novel sequence into the appropriate database to evaluate whether anyone else has described your protein

Suggested Reading

In this chapter we introduced dozens of BLAST servers. In most cases the websites described in Tables 5.1–5.3 contain documentation online.

PSI-BLAST was introduced in an excellent paper by Altschul and colleagues (1997) (see also Suggested Readings for Chapter 4). Further modifications of PSI-BLAST are introduced by Schaffer et al., (2001). The PHI-BLAST algorithm is described by Zhang (1998).

References

Altschul, S. F., et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).

del Portillo, H. A., et al. A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* **410,** 839–842 (2001).

Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* **8,** 967–974 (1998).

Friedberg, I., Kaplan, T., and Margalit, H. Evaluation of PSI-BLAST alignment accuracy in comparison to structural alignments. *Protein Sci.* **9,** 2278–2284 (2000).

Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).

Leonardo da Vinci. *Il Codice Atlantico.* Giunti, Florence, 2000.

Ning, Z., Cox, A. J., and Mullikin, J. C. SSAHA: A fast search method for large DNA databases. *Genome Res.* **11,** 1725–1729 (2001).

Park, J., et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284,** 1201–1210 (1998).

Pearson, W. R., and Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448 (1988).

Schaffer, A. A., et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29,** 2994–3005 (2001).

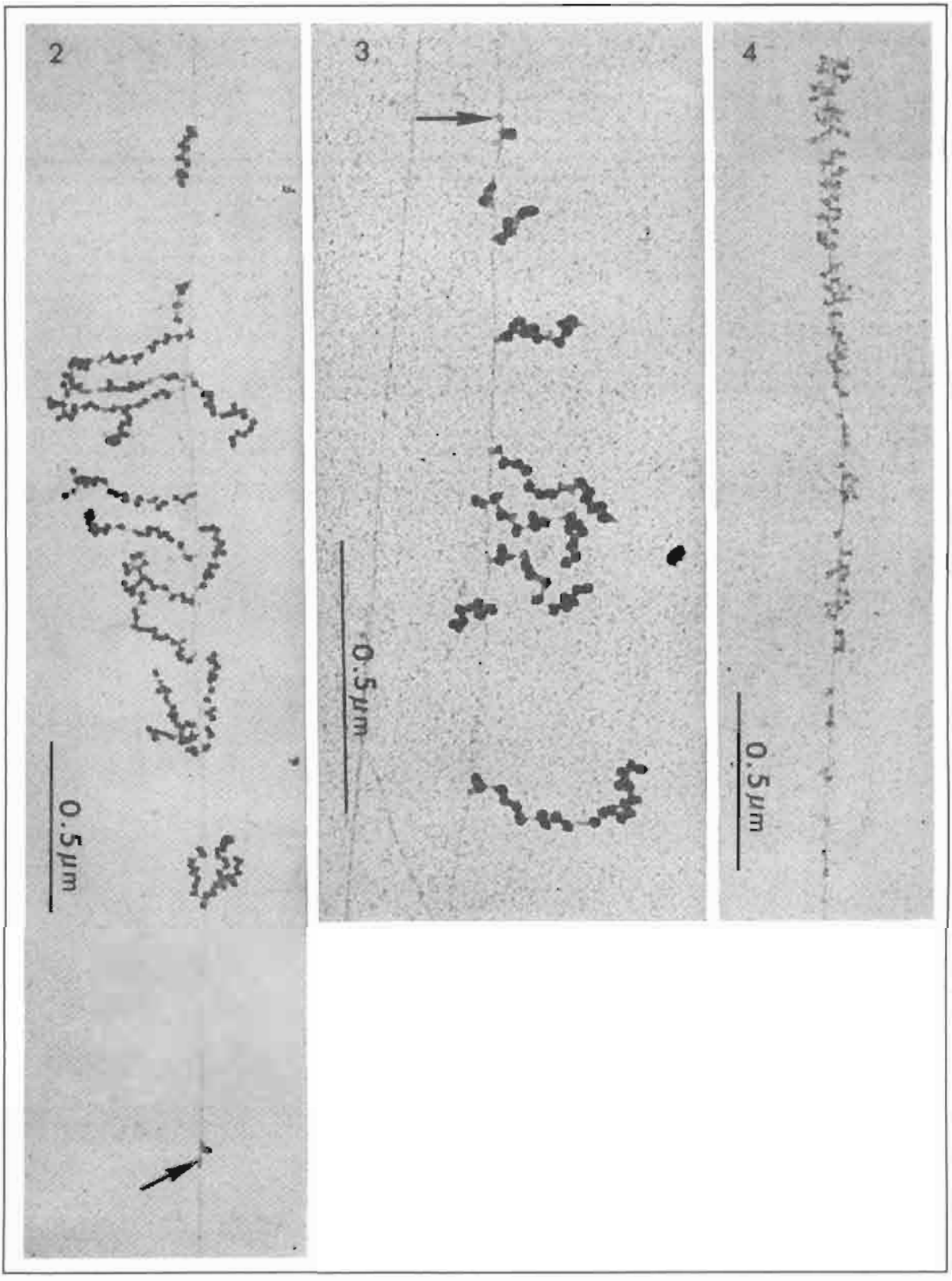
Wootton, J. C., and Federhen, S. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266,** 554–571 (1996).

Zhang, Z., et al. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* **26,** 3986–3990 (1998).

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* **7,** 203–214 (2000).

Part II

Genomewide Analysis of RNA and Protein



Miller and colleagues (1970, p. 394) visualized gene expression. They showed Escherichia coli *chromosomal DNA (oriented vertically as a thin strand in each figure) in the process of transcription and translation. As mRNA is transcribed from the genomic DNA and extends off to the side, polyribosomes (dark objects) appear like beads on a string, translating the mRNA to protein. Used with permission.*

6

Bioinformatic Approaches to Gene Expression

INTRODUCTION

Gene expression occurs when DNA is transcribed into RNA. Each eukaryotic cell contains a nucleus with some 5000–60,000 protein-coding genes, but at any given time the cell expresses only a subset of those genes as mRNA transcripts. The set of genes expressed by a genome is sometimes called the transcriptome. Gene expression is typically regulated in several basic ways:

For the range of gene content in eukaryotic genomes see Chapter 16.

- By region (e.g., brain vs. kidney)
- In development (e.g., fetal vs. adult tissue)
- In dynamic response to environmental signals (e.g., immediate–early response genes that are activated by a drug)
- In disease states
- By gene activity (e.g., mutant vs. wild-type bacterium).

In addition to viewing gene expression as a dynamically regulated process, we can also view proteins and metabolites as regulated dynamically in every cell. See Chapter 8.

The comparison of gene expression profiles has been used to address a variety of biological questions in an assortment of organisms (Fig. 6.1). For viruses and bacteria, studies have focused both on viral and bacterial gene expression and also on the host response to pathogenic invasion. Among eukaryotes, gene expression

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

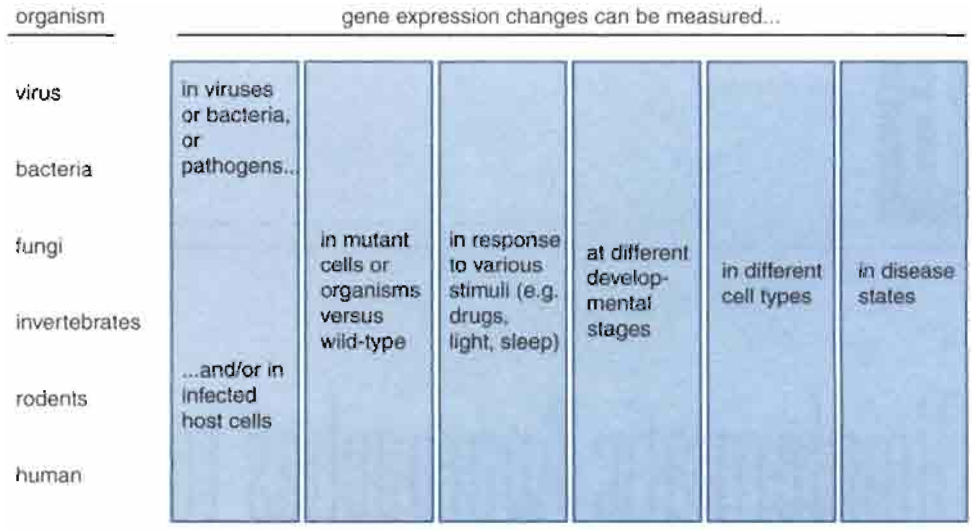


FIGURE 6.1. Gene expression studies are used to address biological questions in a variety of organisms and in a variety of conditions. High-density cDNA or oligonucleotide microarrays have emerged as a powerful tool to examine genomewide changes in gene expression and are described in this chapter and Chapter 7. Other high-throughput techniques are used to measure gene expression changes, such as analysis of expressed sequence tags and serial analysis of gene expression.

studies and in particular microarrays have been employed to address fundamental questions such as the identification of genes activated during the cell cycle or throughout development. In multicellular animals cell-specific gene expression has been investigated, and the effect of disease on gene expression has been studied in rodents and primates, including humans. In recent years, gene expression profiling has become especially important in the annotation of genomic DNA sequences. When the genome of an organism is sequenced, one of the most fundamental issues is to determine which genes it encodes (Chapter 12). Large-scale sequencing of expressed genes, such as those isolated from cDNA libraries (described in this chapter), is invaluable in helping to identify gene sequences in genomic DNA.

In recent decades, gene expression has been studied using a variety of techniques such as Northern blotting, the polymerase chain reaction with reverse transcription (RT-PCR), and the RNase protection assay. Each of these approaches is used to study one transcript at a time. In Northern blotting, RNA is isolated, electrophoresed on an agarose gel, and probed with a radioactive cDNA derived from an individual gene. RT-PCR employs specific oligonucleotide primers to exponentially amplify specific transcripts as cDNA products. RNase protection is used to quantitate the amount of an RNA transcript in a sample based upon the ability of a specific in vitro transcribed cDNA to protect an endogenous transcript from degradation with a ribonuclease. Gene expression may be compared in several experimental conditions (such as normal vs. diseased tissue, cell lines with or without drug treatment). The signals are typically fluorescent or radioactive and may be quantitated. Signals are also normalized to a number of housekeeping genes or other controls that are expected to remain unchanged in their expression levels.

In contrast to these approaches, several high-throughput techniques have emerged that allow a broad survey of gene expression. A global approach to gene expression offers two important advantages over the study of the expression of individual genes:

- A broad survey may identify individual genes that are expressed in a dramatic fashion in some biological state. For example, global comparisons of gene expression in assorted human tissues can reveal which individual transcripts are expressed in a region-specific manner. Milner and Sutcliffe (1983) performed

191 Northern blots, although this experimental approach is not normally employed to measure the expression of so many genes in one study. They found that 30% of the genes were expressed in brain but not in liver or kidney. This type of approach can now be repeated on a larger scale and far more easily with high-throughput techniques that are described in this chapter.

- High-throughput analyses of gene expression can reveal patterns or signatures of gene expression that occur in biological samples. This may include the coordinate expression of genes whose protein products are functionally related. We will examine tools for the analysis of gene expression data (such as clustering trees) in Chapter 7.

Several high-throughput approaches to gene expression are displayed in Figure 6.2. In each case, total RNA or mRNA is isolated from two (or more) biological samples that are compared. The RNA is typically converted to cDNA using reverse transcriptase. Complementary DNA is inherently less susceptible to proteolytic or chemical degradation than RNA, and cDNA can readily be cloned, propagated, and sequenced. In this chapter we will explore three computer-based approaches to the analysis of gene expression: the comparison of cDNA libraries in UniGene, the comparison of serial analysis of gene expression (SAGE) libraries at NCBI, and DNA microarrays.

The enzyme reverse transcriptase, often present in retroviruses, is an RNA-dependent DNA polymerase (i.e., it converts RNA to DNA).

There are other technologies available for the measurement of gene expression, such as differential display and subtractive hybridization. While these approaches have been technically successful for many gene expression problems, they differ

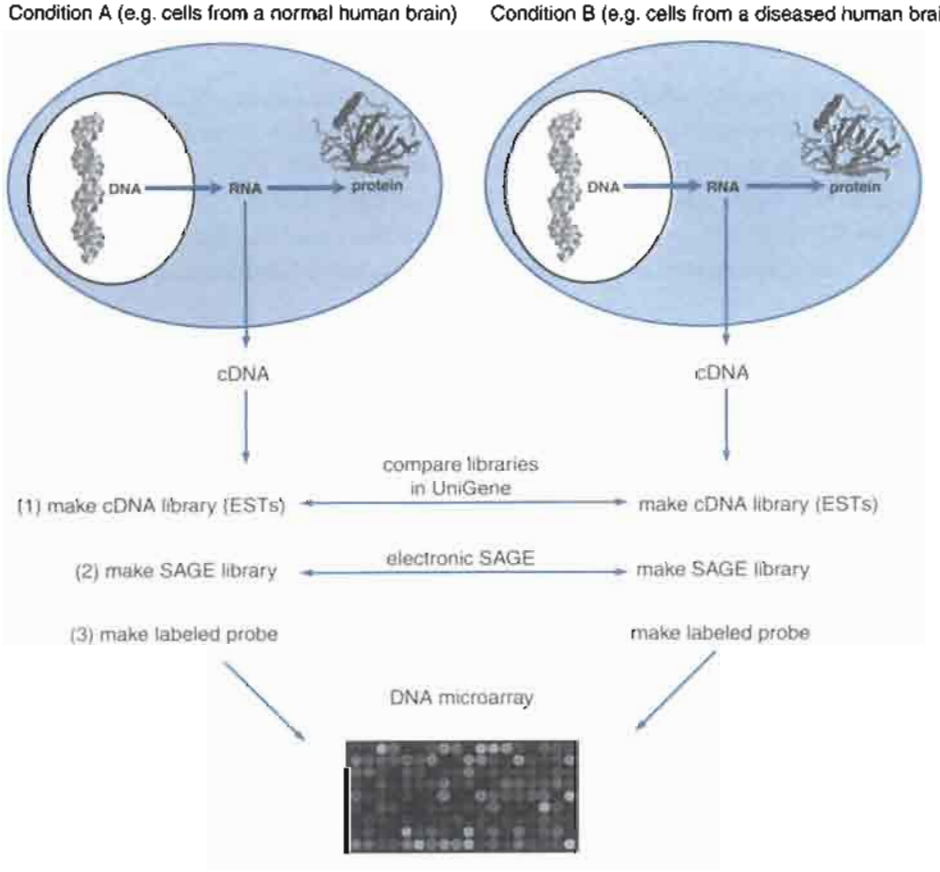


FIGURE 6.2. Gene expression can be measured with a variety of high-throughput technologies. In most cases, two biological samples are compared such as a cell line with or without drug treatment, cells with or without viral infection, or aged versus neonatal rat brain. RNA can be converted to cDNA allowing broader surveys of transcription in a cell. In this chapter and the next we will examine three approaches to gene expression. (1) cDNA libraries can be constructed, generating expressed sequence tags (ESTs). These can be electronically compared in UniGene. (2) Serial analysis of gene expression (SAGE) is another technology in which the abundance of transcripts can be compared. This can also be studied electronically. (3) Complex cDNA mixtures can be labeled with radioactivity or fluorescence and hybridized on DNA microarrays, which contain cDNA or oligonucleotide fragments corresponding to thousands of genes.

from the techniques shown in Figure 6.2 because they generally do not involve the establishment of electronic databases. Techniques used to measure gene expression have been reviewed by Watson and Margulies (1993), Sagerstrom et al. (1997), Vietor and Huber (1997), and Carulli et al. (1998).

While databases of gene expression have been established, it is important to contrast them with DNA databases. A DNA database such as GenBank contains information about the sequence of DNA fragments, ranging in size from small clones to entire chromosomes or entire genomes. The error rate involved in genomic DNA sequencing can be measured (see Chapter 12), and independent laboratories can further confirm the quality of DNA sequence data. In general, DNA sequence does not change for an individual organism across time or in different body regions. In contrast, gene expression is context dependent. A database of gene expression contains some quantitative measurement of the expression level of a specified gene. If two laboratories attempt to describe the expression level of RBP from a cell line, the measurement may vary based on many variables such as the source of the cell line (e.g., liver or kidney), the cell-culturing conditions (e.g., cells grown to subconfluent or confluent levels), the cellular environment (e.g., choice of growth media), the age of the cells, the type of RNA that is studied (total RNA vs. mRNA, each with varying amounts of contaminating biomaterials), the measurement technique, and the approach to statistical analysis. Thus, while it has been possible to create a project such as RefSeq to identify high-quality representative DNA sequences of genes, any similar attempt to describe a standard expression profile for genes must account for many variables related to the context in which transcription occurs.

Ribosomal RNA molecules form structural and functional components of ribosomes, the subcellular units responsible for protein synthesis. Transfer RNA molecules carry a specific amino acid and match it to its corresponding codon on an mRNA during protein synthesis. Transfer RNA consists of about 70–90 nucleotides folded into a characteristic cloverleaf structure. We will discuss databases of RNA sequences and software for finding RNA genes in eukaryotic DNA (Chapter 16).

Richard J. Roberts and Phillip A. Sharp received the 1993 Nobel Prize in Physiology or Medicine for their discovery of "split genes." See ▸http://www.nobel.se/medicine/laureates/1993/.

A molecule in *Drosophila* provides an extraordinary example of alternative splicing. The Down syndrome cell adhesion molecule (DSCAM) gene product potentially exists in more than 38,000 distinct isoforms (Schmucker et al., 2000; Celotto and Graveley, 2001). The gene contains 95 alternative exons that are organized into clusters. Functionally, multiple DSCAM proteins may confer specificity to neuronal connections in *Drosophila*.

mRNA: Subject of Gene Expression Studies

We will now consider what is measured in gene expression studies. In almost every case, total RNA is isolated from cells of interest. This RNA is readily purified using chaotropic agents that separate RNA from DNA, protein, lipids, and other cellular components. Total RNA consists of ribosomal RNA (rRNA; >70% of total RNA), transfer RNA (tRNA; >15%), and messenger RNA (mRNA; about 3–4%). It is the mRNA that is translated into protein. We will consider additional types of RNA in Chapter 12.

Gene expression is regulated in a set of complex steps that can be divided into four categories: transcription, RNA processing, mRNA export, and RNA surveillance (Maniatis and Reed, 2002) (Fig. 6.3).

1. *Transcription.* Genomic DNA is transcribed into RNA in a set of highly regulated steps. In the 1970s, sequence analysis of genomic DNA revealed that portions of the DNA (called exons) match the contiguous open reading frame of the corresponding mRNA, while other regions of genomic DNA (introns) represent intervening sequences that are not present in mature mRNA.

2. *RNA Processing.* Introns are excised from pre-mRNA by the spliceosome, a complex of five stable small nuclear RNAs (snRNAs) and over 70 proteins. Alternative splicing occurs when the spliceosome selectively includes or excludes particular exons (Modrek and Lee, 2002). Pre-mRNA also is capped at the 5′ end. (Eukaryotic mRNAs contain an inverted guanosine called a cap.) Mature mRNA has the unique property among nucleic acids of having a long string of adenine residues attached

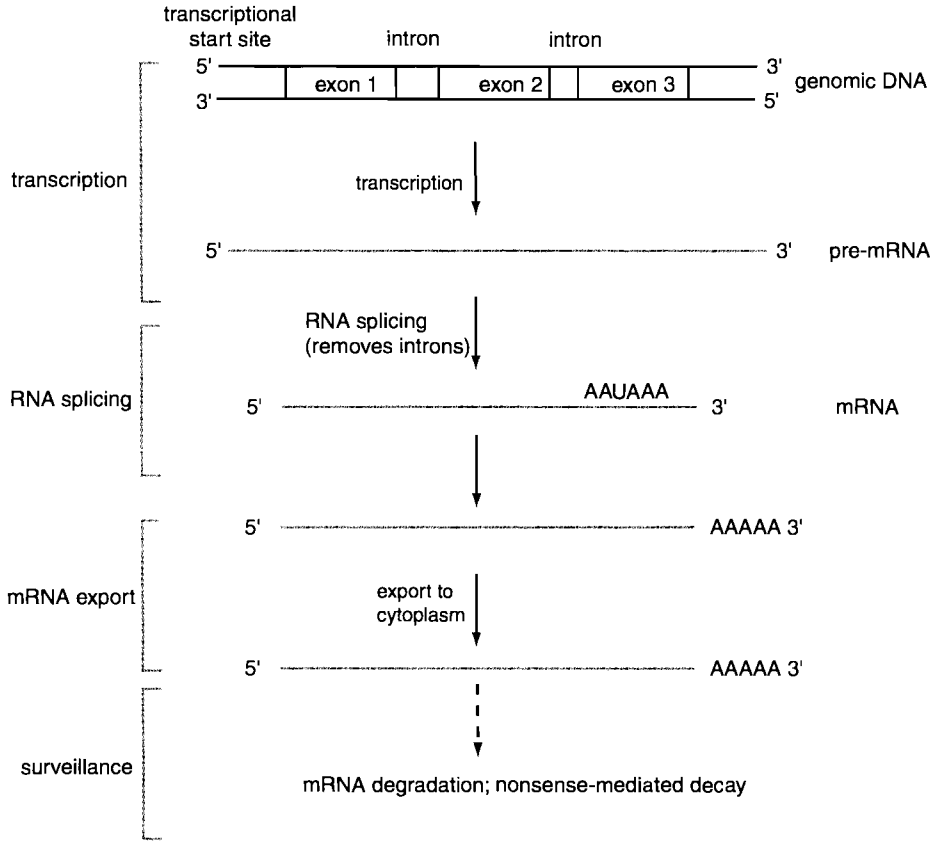


FIGURE 6.3. *RNA processing of eukaryotic genes. Genomic DNA contains exons (corresponding to the mature mRNA) and introns (intervening sequences). After DNA is transcribed, pre-mRNA is capped at the 5′ end, and splicing removes the introns. A polyadenylation signal (most commonly AAUAAA) is recognized, the RNA is cleaved by an endonuclease about 10–35 nucleotides downstream, and a polyA polymerase adds a polyA tail (typically 100–300 residues in length). Polyadenylated mRNA is exported to the cytoplasm where it is translated on ribosomes into protein. An RNA surveillance system involving nonsense-mediated decay degrades aberrant mRNAs; a dashed line indicates that RNA surveillance machinery can also degrade pre-mRNAs.*

to its 3′ end. This tract is typically preceded by the polyadenylation signal AAUAAA or AUUAAA, located 10–35 nucleotides upstream. Polyadenylation of mRNA is extremely convenient from an experimental point of view, because an oligonucleotide (consisting of a string of thymidine residues attached to a solid support [oligo(dT) resin]) can be used to rapidly isolate mRNA to a high degree of purity. In some cases gene expression studies employ total RNA, while many others employ mRNA.

3. *RNA Export.* After splicing occurs, RNA is exported from the nucleus to the cytoplasm where translation occurs.

4. *RNA Surveillance.* An extensive RNA surveillance process allows eukaryotic cells to scan pre-mRNA and mRNA molecules for nonsense mutations (inappropriate stop codons) or frame-shift mutations (Maquat, 2002). This nonsense-mediated decay mechanism is important in the maintenance of functional mRNA molecules. Additional mechanisms control the half-life of mRNAs, targeting them for degradation and thus regulating their availability.

A practical example of how an mRNA sequence aligns with genomic DNA is shown in Fig. 6.4. Go to LocusLink and select the accession numbers for the RBP4 mRNA (i.e., the DNA sequence that corresponds to the mRNA; NM_006744) and a genomic DNA fragment of over 20 million bp on human chromosome 10 that includes the RBP4 gene (NT_030059). A BLAST 2 Sequences pairwise alignment shows an excellent match between the two, with large gaps in the alignment that correspond to introns in the genomic DNA.

It is notable that all of the alignments in Figure 6.4 have gaps, mismatches, or both. These discrepancies reflect polymorphisms or errors associated with either the sequencing of genomic DNA or cDNA. One way to decide which sequence has an error is to look for consistency: If multiple, independently derived genomic DNA clones or expressed sequence tags have the identical nucleotide sequence in a region of interest, you can be more confident that sequence is correct. See Chapter 12 for a further discussion.

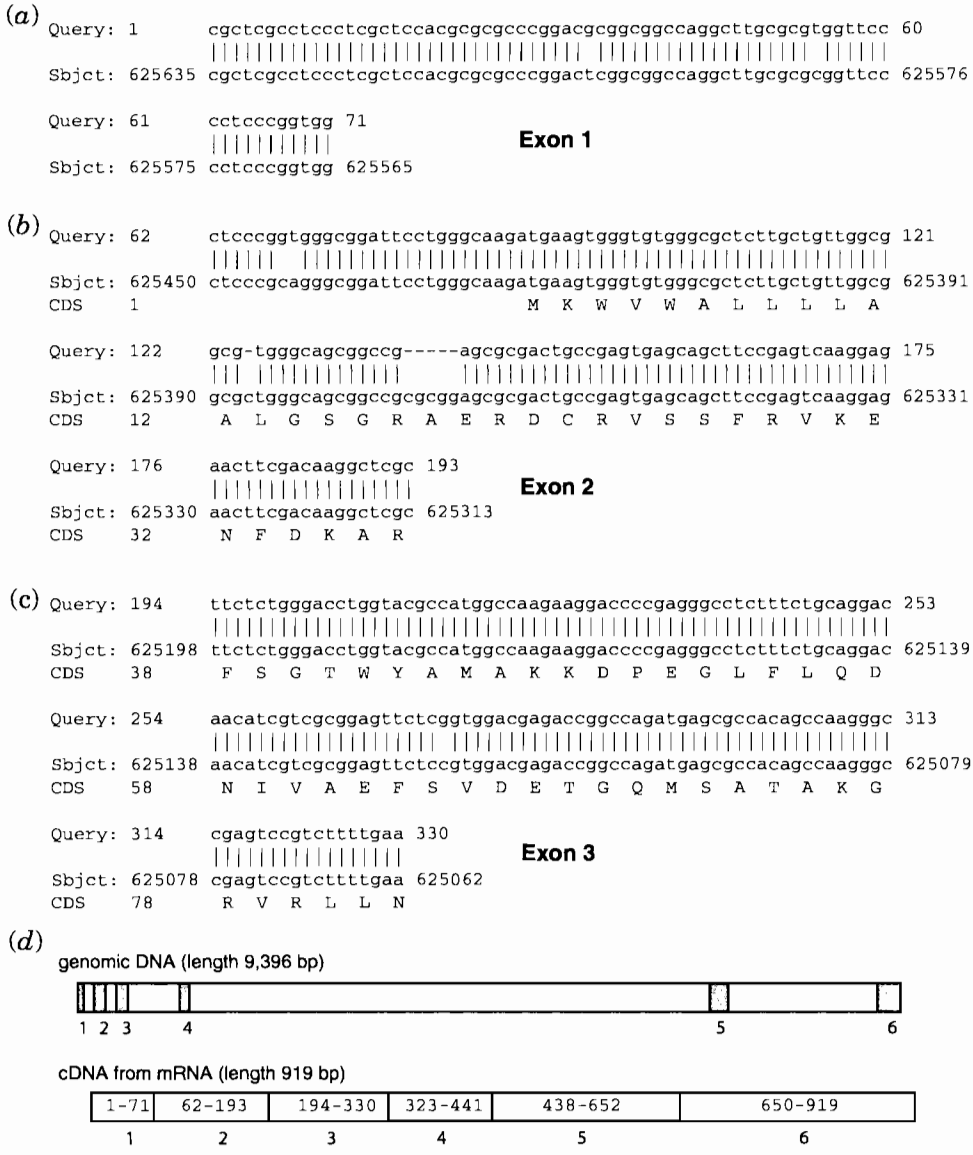


FIGURE 6.4. *Relationship of mRNA to genomic DNA is demonstrated by performing a pairwise sequence alignment between a query consisting of the mRNA-derived sequence for human RBP4 (from LocusLink; NM_006744) and the genomic DNA containing human RBP4 (also from LocusLink; NT_030059). A portion of the alignment from BLAST 2 Sequences is shown in (a)–(c). The genomic DNA fragment, called a contig, consists of 20,700,699 bp from chromosome 10. It aligns with the RBP4 in the bottom-strand orientation (thus the numbers on the subject lines are decreasing). You can follow the sequence of the query from panels as it aligns with portions of the genomic DNA that are interrupted by introns. Based on the alignment, a map showing genomic DNA (top) and the contiguous mRNA (bottom) is shown in (d) by plotting the positions in the pairwise alignment. The exons are numbered 1–6. Note that the scales are different.*

Analysis of Gene Expression in cDNA Libraries

How can we study the majority of mRNA molecules that are expressed from a tissue sample or other biological system of interest? It is technically straightforward to isolate RNA and/or mRNA from a small tissue sample in reasonably large quantities (e.g., hundreds of micrograms of total RNA). However, RNA is both unstable and

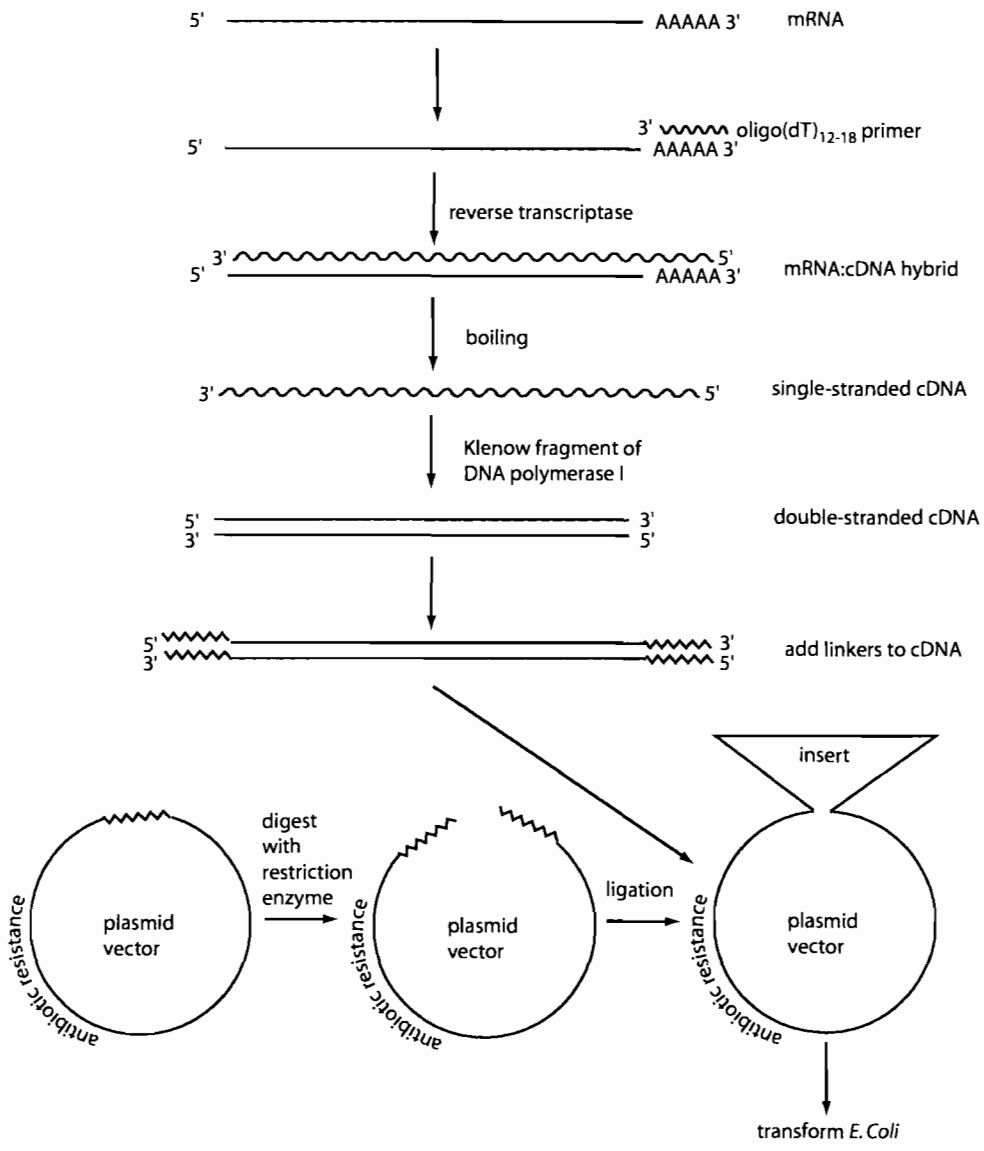


FIGURE 6.5. *Construction of a cDNA library. Messenger RNA is hybridized to an oligo(dT) primer at its 3' polyadenylated tail, and an mRNA:cDNA hybrid is generated by reverse transcription. After boiling to denature the RNA, cDNA is made double stranded with a DNA polymerase. Linkers (e.g., nucleotides recognized by a restriction endonuclease) are added to cDNA so that after appropriate digestion of both the cDNA (also termed the insert) and a plasmid or bacteriophage (also called the vector), the two can be ligated.* Esherichia coli *bacteria are then transformed and selected for antibiotic resistance. In this way, a cDNA library is formed, containing up to thousands of unique cDNA inserts derived from the starting mRNA population.*

highly complex, typically containing thousands of distinct transcripts. One way to solve this problem is to generate a cDNA library (Fig. 6.5). In brief, RNA is converted to double-stranded cDNA, cloned into a vector, and propagated in a bacterial cell line. A vector such as a plasmid has the properties of small size, rapid growth, and the ability to contain a single cDNA insert derived from the starting tissue sample or other biological source. Thousands of cDNA libraries are available commercially; each is derived from a particular organism, cell type, developmental stage, and physiological condition. The clones in a cDNA library may be plated onto Petri dishes. The cDNA inserts, called expressed sequence tags (ESTs), may then be sequenced.

A summary of the number of ESTs in GenBank is available at ▸http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html.

Millions of ESTs have been sequenced, usually as a *single-pass read* of approximately 500 bp from the 3' end and/or the 5' end of the cDNA clone. Adams et al. (1991–1993) pioneered the approach of sequencing thousands of ESTs to identify genes expressed in a particular tissue. (This is called a *shotgun single-pass* approach.) These studies revealed which genes are expressed at the highest relative levels (such as β-actin and myelin basic protein in human brain), and they also described the regional variation in gene expression across different brain regions.

UniGene is accessed via ▸http://www.ncbi.nlm.nih.gov/UniGene/.

We previously described UniGene, a system for partitioning ESTs into a nonredundant set of clusters (Chapter 2). In principle, each unique gene ("unigene") is assigned a single UniGene entry. UniGene encompasses both well-characterized genes and those inferred by the existence of ESTs; all ESTs corresponding to a gene are assigned to that particular UniGene accession number. UniGene clusters containing ESTs that are similar to a known gene are categorized as "highly similar" to that gene (defined as >90% identity in the aligned region), "moderately similar" (70–90% identity), or "weakly similar" (<70% identity).

The number of human UniGene clusters provides one estimate for the number of human genes, although UniGene clusters are retired when two clusters can be joined into one. Each cluster has some number of sequences associated with it, from one (*singletons*) to over 10,000 (Table 6.1). Of the 104,170 clusters in Table 6.1, over one-third are singletons, suggesting that these may be genes expressed so rarely that they have been observed only one time. Alternatively, it is possible that these are not authentic expressed genes but represent some form of library contamination. Indeed, the presence of over 100,000 UniGene clusters is inconsistent with the estimate of some 35,000 protein-coding genes in the human genome (Chapter 17). These singletons may also represent portions of genes that have not yet been sequenced (see Chapter 2).

There are currently 813 ESTs associated with the actin UniGene entry (Hs.1288), and there are 7237 sequences associated with α-tubulin (Hs.334842) (October 2002).

TABLE 6-1 Histogram of Cluster Sizes for Human Entries in UniGene (Build 153, *Homo sapiens*)

Cluster size	Number of clusters
1	36,206
2	14,384
3–4	15,804
5–8	10,612
9–16	5,852
17–32	3,986
33–64	3,516
65–128	4,095
129–256	3,953
257–512	2,170
513–1024	709
1,025–2,048	213
2,049–4,096	70
4,097–8,192	26
8,193–16,384	5
16,385–32,768	1

Source: ▸http://www.ncbi.nlm.nih.gov/UniGene/Hs.stats.shtml.

The presence of thousands of UniGene entries with small cluster sizes suggests that some genes are expressed only rarely. Other genes (such as actin and tubulin) are expressed at very high levels. Even some EST clusters that do not correspond to known, annotated genes are highly represented. For all organisms represented in UniGene, the largest cluster sizes are described in Table 6.2.

DNA hybridization studies from the 1970s suggest that a typical mammalian cell expresses about 300,000 mRNA transcripts, expressed from between 10,000 and 30,000 distinct genes (Hastie and Bishop, 1976). Bishop and colleagues (1974) grouped mRNA into three classes based on relative abundance: (1) genes that are expressed at highly abundant levels (accounting for 10% of the overall transcripts), (2) medium-abundance genes (45% of the mRNA), and (3) low-abundance genes (45% of the mRNA). These three classes correspond to cluster sizes in UniGene (Table 6.1), although the cluster sizes are not formally labeled.

Since all ESTs are derived from a specific region of the body at a particular time of development and a particular physiological state, there is inherently a large amount of information associated with the analysis of many ESTs (Schuler, 1997). There are two main approaches to extracting information from UniGene. First, if we want to know where in the body a particular gene (such as RBP) is expressed, we can survey UniGene. The number of ESTs associated with that gene reflects the abundance of the transcript (but see the discussion below), and the tissue source of the libraries from which ESTs are derived reflects the regional distribution of ESTs. This approach is sometimes called an *electronic Northern blot.*

A second approach to extracting information from UniGene is to electronically subtract cDNA libraries. Electronic cDNA library subtraction in UniGene also

TABLE 6-2 Ten Largest Cluster Sizes in UniGene (October 2002).

To obtain these data, all UniGene files were downloaded by file transfer protocol (FTP) and sorted by size in a relational database (DRAGON) (Bouton and Pevsner, 2000).

UniGene Identifier	Species	Cluster Size	Gene Name
Hs.181165	*H. sapiens*	25,232	Eukaryotic translation elongation factor 1 alpha 1
Hs.169476	*H. sapiens*	14,277	Glyceraldehyde-3-phosphate dehydrogenase
Ta.9227	*Triticum aestivum*	14,231	Ubiquitin
Hs.14376	*H. sapiens*	12,749	Actin, gamma 1
Mm.196614	*Mus musculus*	10,649	Eukaryotic translation elongation factor 1 alpha 1
Hs.356360	*H. sapiens*	10,596	Ribosomal protein S2
Mm.30266	*M. musculus*	10,290	Hemoglobin, beta adult major chain
Hs.356428	*H. sapiens*	9,987	mRNA expressed only in placental villi
Hs.288061	*H. sapiens*	9,667	Actin, beta
Dr.2984	*Danio rerio*	9,058	40S ribosomal protein S18

allows the cDNA sequences in two populations to be compared using the digital differential display tool (Fig. 6.6*a*).

You can access this tool from the main UniGene page by selecting an organism (e.g., *H. sapiens*), then clicking "Digital Differential Display" (Fig. 6.6*b*). A comparison of muscle libraries to brain libraries reveals many transcripts selectively associated with muscle (e.g., myoglobin, troponins) (Fig. 6.7) and others selectively associated with brain. The output shows a dot for each gene whose intensity corresponds to the expression level of that gene. A probability value is associated with each transcript using a Fisher's exact test. This is used to test the null hypothesis that the number of sequences for any given gene in the two pools (e.g., myosin) is the same in either pool (Table 6.3).

The p value for a Fisher's exact test is given by

$$p = \frac{N_A!\, N_B!\, c!\, C!}{(N_A + N_B)!\, g1_A!\, g1_B!\, (N_A - g1_A)!\, (N_B - g1_B)!} \qquad (6.1)$$

The null hypothesis (that gene 1 is not differentially regulated between brain and muscle) is rejected when the probability value p is less than $0.05/G$, where 0.05 is the nominal threshold for declaring significance and G is the number of UniGene clusters analyzed (thus, G is a conservative Bonferroni correction).

The comparison of gene expression profiles using databases of libraries may be considered a tool that rapidly provides candidate genes for further analysis. In our muscle-versus-brain library comparison, some of the regulated genes identified with digital differential display correspond to "hypothetical proteins" that have not been functionally characterized. These could be studied in further detail. Schmitt and colleagues (1999) used a similar approach to identify 139 transcripts that are selectively up regulated in breast cancer tissue. Such transcripts could provide markers for the early detection of breast cancer or they could reflect changes in tumor

Experimentally, the differential display technique allows two RNA (or corresponding cDNA) sources to be subtracted from each other. One population is labeled selectively (e.g., by ligating an oligonucleotide sequence to the ends of one population of cDNAs), the two populations are hybridized to form duplexes between clones shared in common in the two populations, and cDNA clones that are strongly overrepresented in one of the two original populations are selectively amplified. These clones are then sequenced.

While the NCBI website employs Fisher's exact test, other statistical approaches to cDNA library comparison have been described. In particular, Stekel et al. (2000) developed a log-likelihood procedure to assess the probability that gene expression differences observed in a comparison of two or even multiple cDNA libraries are due to genuine transcriptional differences and not sampling errors.

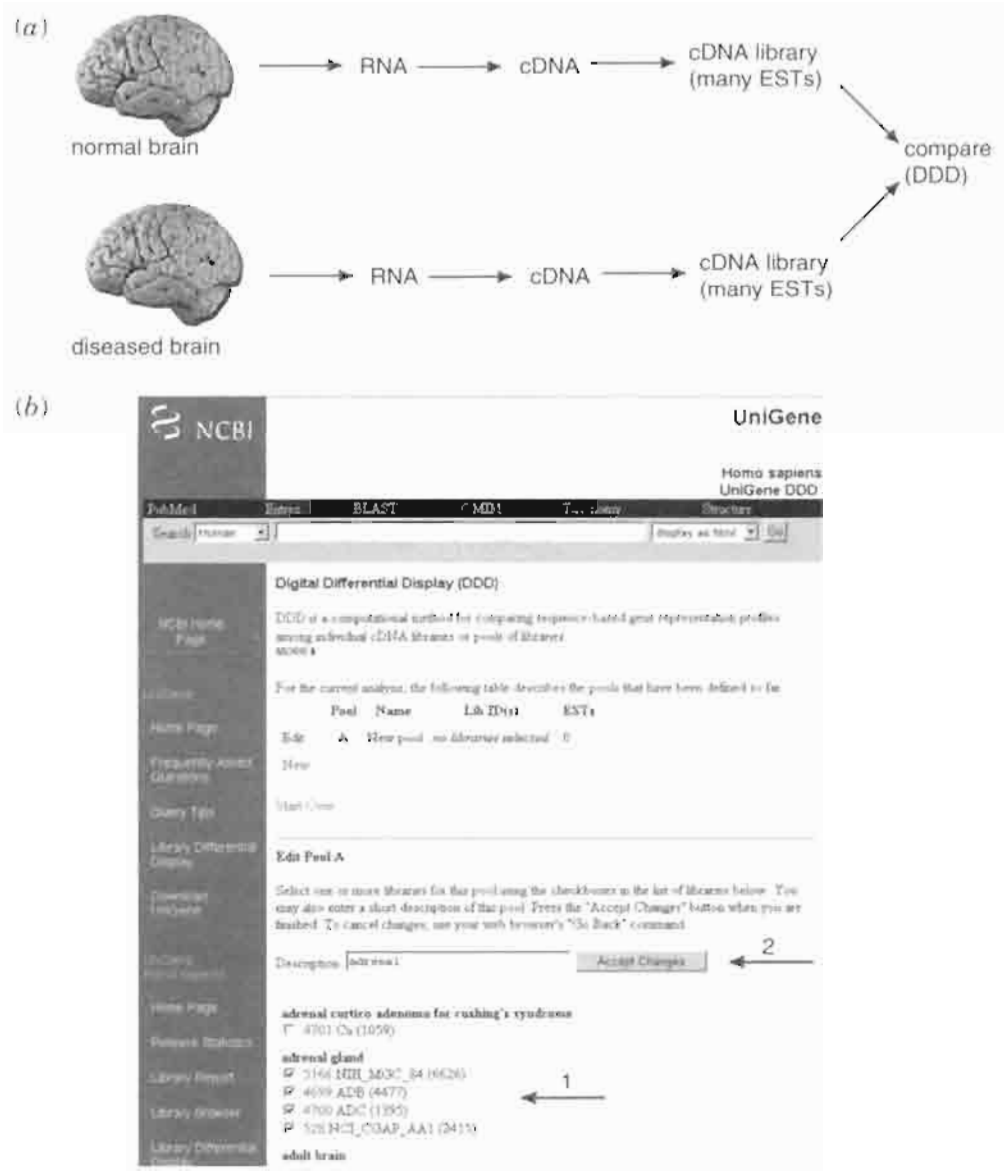


FIGURE 6.6. Digital differential display (DDD) is used to compare the content of expressed sequence tags (ESTs) in cDNA libraries from UniGene. Over 1000 libraries have been generated by isolating RNA from a tissue source [such as a normal versus diseased human brain in (a)], synthesizing cDNA, packaging the cDNA in a cDNA library, and sequencing up to thousands of cDNA clones (ESTs) from each library. (b) The clones in each library (or in pools of libraries) may be compared using DDD. This site is accessed from the NCBI UniGene site; on the left sidebar click Homo sapiens, *then select "Library digital differential display." At this site, click boxes corresponding to any library (or set of libraries; arrow 1) then press "Accept changes" (arrow 2). You will then be given the opportunity to select a second library (or second set of libraries) for comparison.*

tissue that offer targets for therapeutic intervention. Many other studies have employed EST sequencing and/or electronic analyses of sequenced cDNA libraries (e.g., Carulli et al., 1998).

Pitfalls in Interpreting Expression Data from cDNA Libraries

The contents of cDNA libraries in UniGene and elsewhere must be analyzed with caution for several reasons:

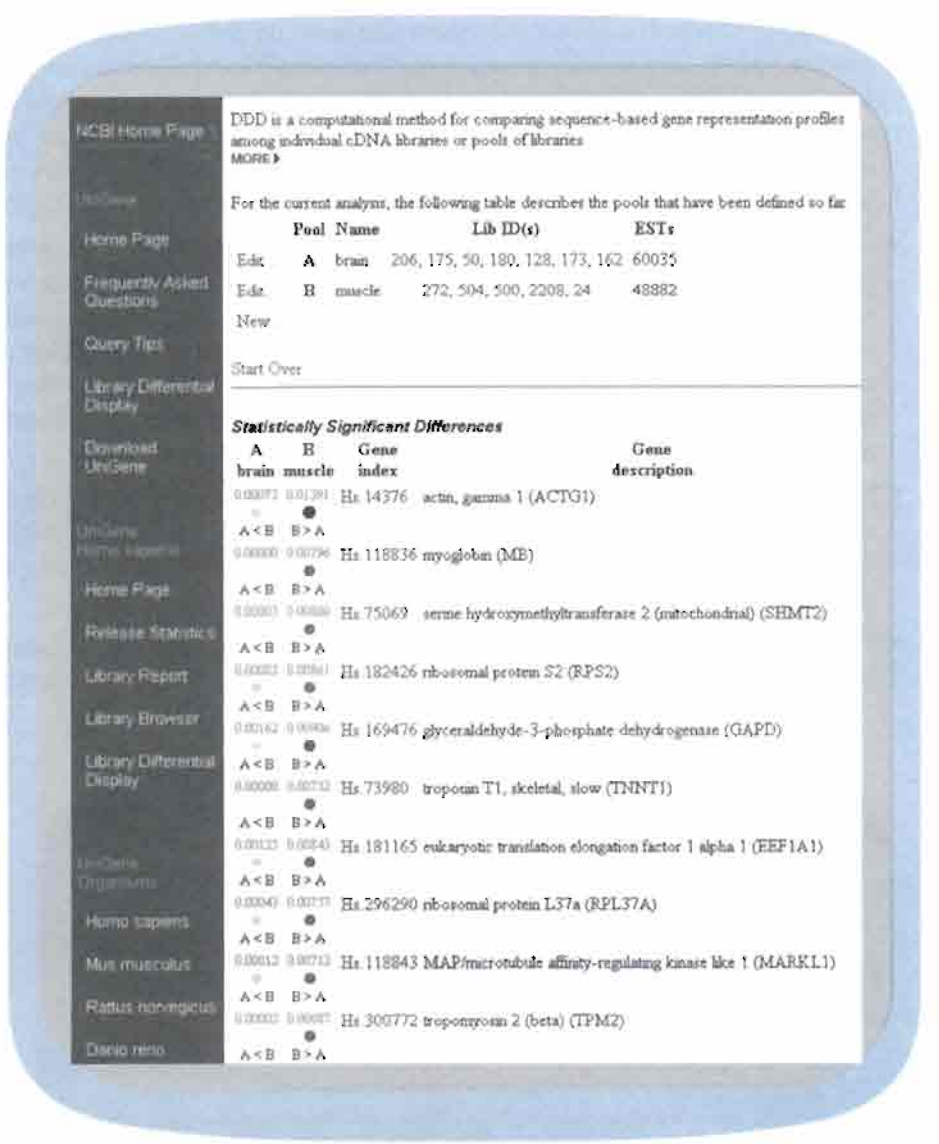


FIGURE 6.7. Result of an electronic comparison of cDNA libraries using the DDD tool of UniGene. The results are displayed as a list of genes (with UniGene accession numbers) for genes that are preferentially expressed in one or the other pool of libraries. Here, a variety of muscle-specific transcripts are displayed (e.g., troponin, tropomyosin). Other transcripts such as a neuroblastoma gene (not shown) are more highly represented in brain-derived libraries.

- Investigators choose which libraries to construct, and there is likely to be bias toward familiar tissues (such as brain and liver) and bias away from more unusual tissues. The rat nose contains over two dozen secretory glands, almost all of which are of unknown function, but for most of these glands cDNA libraries have never been constructed.

TABLE 6-3 Fisher's 2 × 2 Exact Test Used to Test Null Hypothesis That A Given Gene (Gene 1) Is Not Differentially Regulated in Two pools

	Gene 1	All Other Genes	Total
Pool A (e.g., brain)	Number of sequences assigned to gene 1 ($g1_A$)	Number of sequences in this pool NOT gene 1 ($N_A - g1_A$)	N_A
Pool B (e.g., muscle)	Number of sequences assigned to gene 1 ($g1_B$)	Number of sequences in this pool NOT gene 1 ($N_B - g1_B$)	N_B
Total	$c = g1_A + g1_B$	$C = (N_A - g1_A) + (N_B - g1_B)$	

Source: Adapted from Claverie (1999) and ▸ http://www.ncbi.nlm.nih.gov/UniGene/fisher.shtml.

In UniGene, click *Homo sapiens*, then "library browser," to see the range of clones that are sequenced in typical libraries.

- The depth to which a library is sequenced affects its ability to represent the contents of the original cell or tissue. A cDNA library is expected to contain a frequency of clones that faithfully reflects the abundance of transcripts in the source material. By sequencing only 500 clones, it is unlikely that many low-abundance transcripts will be represented when the contents of the entire library are analyzed. In practice, cDNA libraries are sequenced to varying depths.
- Another source of bias is in library construction. Many libraries are normalized, a process in which abundant transcripts become relatively underrepresented while rare transcripts are represented more frequently. The goal in normalizing a library is to minimize the redundant sequencing of highly expressed genes and to thus discover rare transcripts (Bonaldo et al., 1996). It would be inappropriate to compare normalized and nonnormalized libraries directly using a tool such as UniGene's differential display.
- ESTs are often sequenced on one strand only, rather than thoroughly sequencing both top and bottom strands. Thus, there is a substantially higher error rate than is found in finished sequence. (We will discuss sequencing error rates in Chapter 12.)
- Chimeric sequences can contaminate cDNA libraries. For example, two unrelated inserts are occasionally cloned into a vector during library construction.

FIGURE 6.8. *The Mammalian Gene Collection (MGC) is a project to sequence a representative open reading frame clone for each human and mouse gene. Currently, a nonredundant set of over 20,000 human and mouse clones have been assembled. Of these clones, 3000 have been fully sequenced, and 75% of the selected clones have full open reading frames. MGC, sponsored by the NIH, will provide a reliable, high-quality clone resource to the research community. The site (*http://mgc.nci.nih.gov/*) is searchable by BLAST.*

TIGR Gene Indices

While our discussion of ESTs and cDNA libraries has focused on UniGene, other important resources are available. The Institute for Genomic Research provides the TIGR Gene Indices, a collection of ESTs organized into several dozen species-specific databases (Quackenbush et al., 2001; Lee et al., 2002). The approach taken by TIGR is to focus on the analysis of EST sequences to assemble unique genes called tentative consensus (TC) sequences. This emphasis on clustering and assembly results in a collection of consensus sequences corresponding to genes. TIGR Gene Indices are then useful for a variety of purposes not as readily available with UniGene. (While UniGene does not assemble ESTs into a single cluster, NCBI does provide a list of the longest EST sequence from each cluster.) The TIGR Eukaryotic Gene Orthologs (EGO) database consists of orthologous genes identified by pairwise alignments of TC sequences, allowing the comparison of homologous genes across dozens of organisms. The TIGR Gene Indices are also accessible by blastn and tblastn (Table 5.1 and Figs. 5.3–5.5).

ESTs can be related to full-length clones several ways. The Integrated Molecular Analysis of Genomes and Their Expression (IMAGE) consortium has collected libraries and made them available for distribution. The Mammalian Gene Collection (MGC) is a NIH project that also offers full-length human and mouse clones. Its site (Fig. 6.8) can be searched by BLAST.

The IMAGE consortium website (▸http://image.llnl.gov/) can be queried for clones from a number of species.

Serial Analysis of Gene Expression (SAGE)

Serial analysis of gene expression allows the quantitative measurement of gene expression by measuring large numbers of transcripts from tissues of interest. Short tags of 9–11 bp of DNA are isolated from the 3′ end of transcripts, sequenced, and assigned to genes.

The procedure for producing SAGE tags is outlined in Figure 6.9 (Velculescu et al., 1995). RNA is isolated from a source of interest and converted to cDNA with a biotinylated oligo(dT) primer. A restriction enzyme (the "anchoring enzyme") is used to digest the total population of transcripts so that only short fragments are isolated, and the tight interaction between biotin and avidin allows the 3' end of each transcript to be tethered to streptavidin beads. Two populations of linkers are added, allowing the cDNA to be digested with a specialized restriction enzyme that releases the linker with a short fragment of cDNA (the "tag"). Tags are concatenated, cloned, and sequenced. This process results in the description of thousands (or millions) of tags from a biological source.

The SAGE site at NCBI is at ▸http://www.ncbi.nlm.nih.gov/SAGE/.

A variety of SAGE libraries have been constructed. Each tag in a library is likely to correspond to a single gene. For a 9-bp tag, there are 4^9, or 262,144, transcripts that can be distinguished, assuming a random nucleotide distribution at the tag site. In practice, tags are mapped to genes using UniGene. In some cases, a tag may be present on more than one gene. In other cases, a gene may have more than one tag (e.g., there may be alternative splicing of a transcript such that there are multiple tags for that gene). An assumption of SAGE is that the number of tags found in a SAGE library is directly proportional to the number of mRNA molecules in that biological sample.

SAGE has been used to describe the properties of the yeast transcriptome (Velculescu et al., 1997). The expression of 4665 genes was characterized, the majority

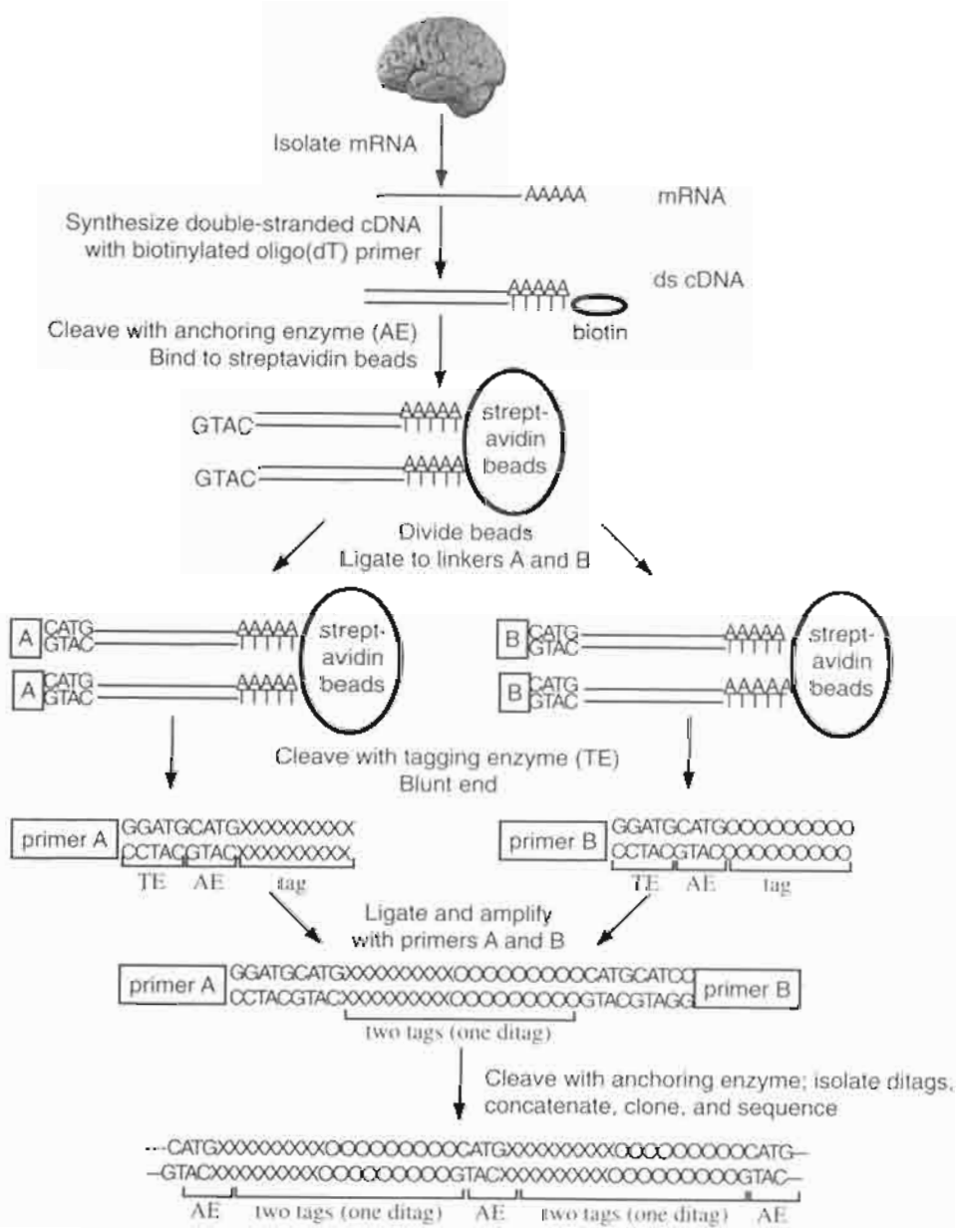


FIGURE 6.9. Description of serial analysis of gene expression (SAGE). Messenger RNA is isolated from a source (such as brain), and double-stranded cDNA is synthesized using a biotinylated oligo(dT) primer. The cDNA is cleaved with a four-cutter restriction endonuclease ("anchoring enzyme," AE) that cleaves most transcripts in a cell. The 3' portion of each transcript is immobilized on streptavidin beads (large ovals), and linkers (A or B) are added containing restriction endonuclease recognition site ("tagging enzyme," TE). Cleavage of the ligated clones with the tagging enzyme releases the linker with the cDNA tags (X's at left, O's at right). The ligated tag pairs are concatenated, cloned, and sequenced. Each tag represents a fragment of 9 bp of a transcript. Modified from Velculescu et al. (1995). Used with permission.

of which had not been functionally characterized. Consistent with the analysis of UniGene clusters (Table 6.1), these data showed that many transcripts are expressed only rarely. Zhang et al. (1997) used SAGE to profile gene expression in normal and neoplastic gastrointestinal tissue. They estimated that the number of distinct transcripts that were expressed in each cell type ranged from about 14,000 to 20,000, and the expression levels ranged from one copy per cell to 5300 copies per cell. The abundance of each gene was estimated by dividing the observed number of tags for a transcript by the total number of tags obtained.

SAGE libraries can be electronically queried at the NCBI website, allowing the comparison of gene expression in any tissues for which SAGE libraries have been generated (Lash et al., 2000; Lal et al., 1999). The website includes tag data from SAGE libraries and annotation data in which tags are mapped to genes (Fig. 6.10). SAGE libraries can be selected in a manner similar to using digital differential

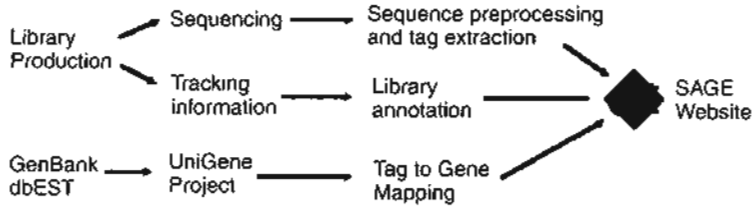


FIGURE 6.10. Build process for the NCBI SAGE database includes a data generation portion (top; library production) and a mapping portion (bottom).

display (Fig. 6.11). The genes that correspond to tags differentially present in lung include surfactant, pronapsim A and secretoglobin with hundreds of tags in lung but none in brain (Fig. 6.12). Assorted brain-enriched transcripts are also identified (Fig. 6.12). Examination of surfactant (by clicking its link) shows that the mapping of this particular tag (TGCCAGGTCT) to the surfactant gene (UniGene Hs.177852) appears unambiguous, and 50 tags corresponding to surfactant have been identified selectively in a lung library (Fig. 6.13).

The "Retrieve by sequence" option accomplishes an electronic version of a conventional Northern blot, in which the expression distribution of an individual gene

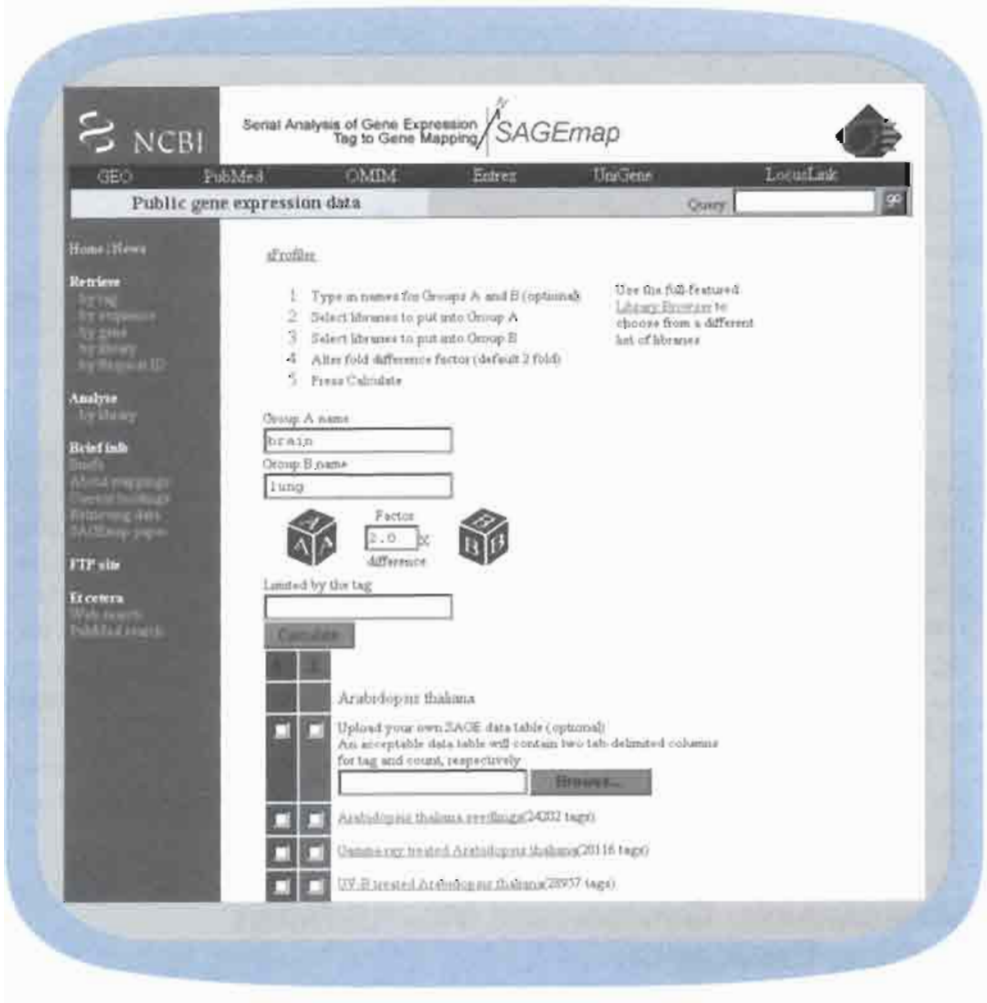


FIGURE 6.11. The NCBI SAGE website library comparison tool (obtained by clicking "Analyze by library" at http://www.ncbi.nlm.nih.gov/SAGE/) allows two pools of SAGE tags to be evaluated in order to compare gene expression profiles. In this case, boxes were clicked under headings A and B that correspond to human brain and lung SAGE libraries.

Tag	Counts (brain) pool A	Counts (lung) pool B	p value	Unigene cluster	Cluster title
AACAGCAAAA	340	1	1.000	Hs.73133	MT3 metallothionein 3 (neurotrophic)
AACAGCAAAA	340	1	1.000	Hs.196550	LOC163590 AF464140
CGCAGCGGGT	0	270	1.000	Hs.322854	NAP1 (pronapsin A)
CAGGAGGAGT	63	1	1.000	Hs.13751	GRP58 glucose regulated protein
CAGGAGGAGT	63	1	1.000	Hs.40411	ESTs, Highly similar to S55507
CAACTAATTC	970	38	1.000	Hs.75106	CLU clusterin
CAACTAATTC	970	38	1.000	Hs.69997	ZNF238 zinc finger protein 238
CTTTGAGTCC	0	228	1.000	Hs.2240	SCGB1A1 secretoglobin (uteroglobin)
TCCCTATAAG	211	0	1.000		No reliable Unigene cluster found
GGCAAGAAAA	0	25	1.000	Hs.405528	RPL27 ribosomal protein L27
TGCCAGGTCT	0	50	1.000	Hs.177582	SFTPA2 surfactant, pulmonary-assoc.
GCGACCGTCA	447	8	1.000	Hs.273415	ALDOA (aldolase A, fructose-bispho.)
AAGAGCGCCG	87	0	1.000	Hs.75452	HSPA1A heat shock 70kDa protein 1A
AAGAGCGCCG	87	0	1.000	Hs.274402	HSPA1B heat shock 70kDa protein 1B
AAGCCAGCCC	87	0	1.000	Hs.1432	PRKCSH protein kinase C substrate
ACTCCCATAT	86	0	1.000	Hs.380027	CROC4 activator of c-fos promoter
TCCCATTAAG	253	1	1.000		No reliable Unigene cluster found
ACGAGGGGTG	85	0	1.000	Hs.6194	BCAN chondroitin sulfate proteoglycan
GTATGGGCCC	247	1	1.000	Hs.75184	CHI3L1 chitinase 3-like 1
GCCTGTCCCT	80	0	1.000	Hs.821	BGN (biglycan)

FIGURE 6.12. *Result of an electronic comparison of SAGE libraries from brain and lung shows SAGE tags corresponding to transcripts that are present in different abundance in the pools. In this example, group B (lung) includes tags corresponding to surfactant and genes known to be expressed preferentially in lung. Additional tags correspond to genes expressed in brain such as a neurotrophic metallothionein.*

USAGE is a web-based relational database that offers comprehensive analyses of SAGE data (▸http://www.cmbi.kun.nl/usage) (van Kampen et al., 2000). The Human Transcriptome Map was developed at the University of Amsterdam (▸http://bioinfo.amc.uva.nl/HTM). Based on the analysis of 2.4 million SAGE tags, Caron et al. (2001) described expression profiles for each human chromosome. They identified dozens of discrete regions of increased gene expression on chromosomal domains.

SAGE is another extremely useful website (▸http://www.sagenet.org/). In addition to offering data, protocols, and SAGE maps, the site also has an extensive bibliography of SAGE publications.

is assessed across various tissues. By entering an accession number corresponding to the DNA sequence of RBP (Fig. 6.14), a corresponding SAGE tag is shown (see arrow 1). Clicking the link on this tag shows the "retrieve by SAGE tag" site (Fig. 6.15). Libraries in which this tag have been identified are listed, as well as the frequency (in tags per million) (Fig. 6.15a). Here, RBP4 is detected most abundantly in a normal liver library. This page also presents a summary of genes that contain this tag (Fig. 6.15b). In this example, the tag maps to four different UniGene clusters, but only the RBP4 cluster is listed as "reliable" as determined by NCBI using criteria such as the availability of corresponding genomic DNA data.

There are other web-based resources that allow the electronic comparison of SAGE libararies, such as USAGE (van Kampen et al., 2000) and the Human Transcriptome Map (Caron et al., 2001).

MICROARRAYS: GENOMEWIDE MEASUREMENT OF GENE EXPRESSION

DNA microarrays have emerged as a powerful technique to measure gene expression. While EST sequencing projects and SAGE allow high-throughput analyses of

FIGURE 6.13. *SAGE tag corresponding to surfactant (see Fig. 6.12) shows that this transcript has been identified exclusively in normal lung. You can obtain this result by clicking "Retrieve by tag" at the NCBI site.*

gene expression, it is microarrays that have been used most broadly to assess differences in mRNA abundance in different biological samples. The use of microarrays has increased rapidly since the pioneering work of Patrick Brown and colleagues at Stanford University, Jeffrey Trent and colleagues at the NIH, and others (De Risi et al., 1996).

A microarray is a solid support (such as a glass microscope slide or a nylon membrane) on which DNA of known sequence is deposited in a regular gridlike array. The DNA may take the form of cDNA or oligonucleotides, although other materials (such as genomic DNA clones; Chapter 18) may be deposited as well. Typically, several nanograms of DNA are immobilized on the surface of an array. RNA is extracted from biological sources of interest, such as cell lines with or without drug treatment, tissues from wild-type or mutant organisms, or samples studied across a time course. The RNA (or mRNA) is often converted to cDNA, labeled with fluorescence or radioactivity, and hybridized to the array. During this hybridization, cDNAs derived from RNA molecules in the biological starting material can hybridize selectively to their corresponding nucleic acids on the microarray surface. Following washing of the microarray, image analysis and data analysis are performed to quantitate the signals that are detected. Through this process, microarray technology allows the simultaneous measurement of the expression levels of thousands of genes represented on the array.

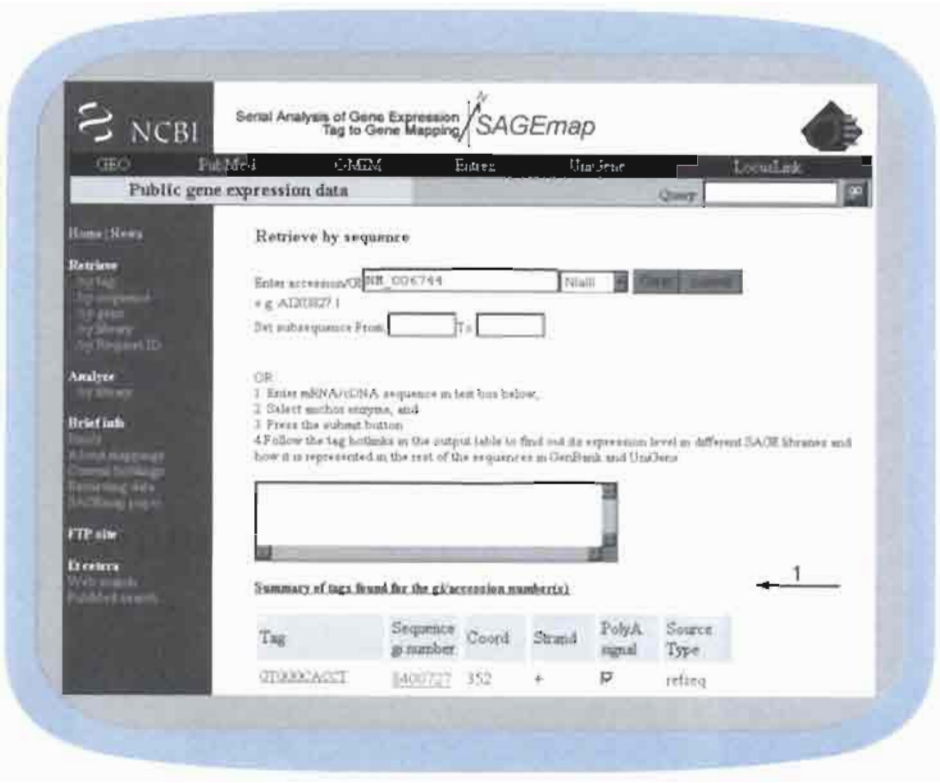


FIGURE 6.14. Retrieve by sequence tool allows the user to assess the distribution and abundance of a transcript in various tissues for which SAGE libraries are available. Here, the accession number for RBP4 is entered (NM_006744). A summary of tags is then shown (arrow 1).

The term *functional genomics* refers to the large-scale analysis of the genomewide function of genes, in contrast to the study of individual protein and nucleic acid molecules. Microarray-based gene expression experiments form the core of functional genomics. There has been enthusiasm for microarrays in the research community because of their potential to yield large amounts of information (Table 6.4). Notably, the rapid accumulation of molecular sequence data in GenBank has led to the availability of many thousands of clones of unknown function. DNA sequences corresponding to both known genes and poorly characterized ESTs have been deposited on microarrays, potentially allowing their function to be determined. The potential of this technology is great, and fundamental biological insights have already been obtained (see Chapter 7).

It is also important to realize the limitations of microarray technology (Table 6.5). The costs of microarrays have caused many investigators to perform relatively few control experiments to assess the reliability and validity of their findings. There are many quality control issues; for example, most researchers cannot assess the nature of the DNA immobilized on an array. Anecdotal reports suggest that a large proportion of DNA sequences are incorrect (e.g., because they are incorrectly annotated or multiple clones are spotted on an array at a supposedly unique address). As the field of microarray technology matures, this error rate is likely to diminish.

The ultimate product of gene expression is not mRNA but protein. The relationship between mRNA levels and protein levels in a cell is poorly characterized. If an mRNA transcript is elevated twofold, is the amount of protein also increased twofold? What is the significance of gene expression changes to a cell? Answers to these questions are only beginning to emerge. At present, high-throughput protein

(*a*)

SAGE library data for this tag :

Library name	Tags per million	Tag counts	Total tags
SAGE Br N mammary gland epithelium ductal normal antibody purified CGAP non-normalized SAGE library method bulk	26	1	37642
SAGE normal pool(6th) brain normal SAGE CGAP non-normalized SAGE library method bulk	63	4	63208
SAGE Duke Kidney kidney normal CGAP non-normalized SAGE library method bulk	95	4	41857
SAGE Caco 2 adenocarcinoma colon SAGE CGAP non-normalized SAGE library method cell line	113	7	61667
SAGE normal liver liver normal CGAP non-normalized SAGE library method bulk	433	29	66861

(*b*)

Summary of genes found for this tag

UniGene cluster id(s)	UniGene cluster title	RefSeq	Number of seqs	contig	MGC	mRNA	ESTs with PA	EST	Get seqs
Hs.418063	RBP4 retinol binding protein 4, plasma	NM_006744	3/6	☑	1	1	0	103	go
Hs.381462	FLJ14251 hypothetical protein FLJ14251		1/6						go
Hs.272366	CYP M cytochrome P450 monooxygenase		1/6						go
Hs.75315	AKR1B1 aldo keto reductase family 1, member B1 (aldose reductase)		1/6						go

FIGURE 6.15. *Result of a virtual Northern blot experiment using RBP4 as a query. Data for the tag (GTGGGCACCT) are shown here. This tag is mapped to RBP4. (a) The result shows that RBP is present seven times in a Caco-2 (colorectal carcinoma) cell line with a frequency of 113 tags per million. When only one or a few tags are identified in a library, as occurs in some cases here, it becomes important to explore whether the results are false positives. SAGE tags can be inappropriately assigned to particular genes or to particular libraries. (b) A summary of genes found for this tag shows that four UniGene clusters share this sequence.*

analyses are technically more difficult to perform (especially protein arrays; Chapter 8) than transcriptional profiling studies.

Several groups have reported a weak positive correlation between mRNA levels and levels of the corresponding proteins in the yeast *Saccharomyces cerevisiae* and other systems (Futcher et al., 1999; Greenbaum et al., 2002). Greenbaum et al.

TABLE 6-4 Major Advantages of Microarray Experiments

Advantage	Comment
Fast	One can obtain data on the expression levels of over 10,000 genes within one week.
Comprehensive	The entire yeast genome can be represented on a chip.
Flexible	cDNAs or oligonucleotides corresponding to any gene can be represented on a chip.

TABLE 6-5 Major Disadvantages of Microarray Experiments

Disadvantage	Comment
Cost	Many researchers find it prohibitively expensive to perform sufficient replicates and other controls.
Unknown significance of RNA	The final product of gene expression is protein, not RNA.
Uncertain quality control	It is impossible for most investigators to assess the identity of DNA immobilized on any microarray. Also, there are many artifacts associated with image analysis and data analysis.

(2002) performed a meta-analysis of gene expression and protein abundance data sets and suggested that there is a broad agreement between mRNA and protein levels.

An overview of the procedures used in a microarray experiment is shown in Fig. 6.16, arbitrarily divided into eight stages. We will consider each of the stages below.

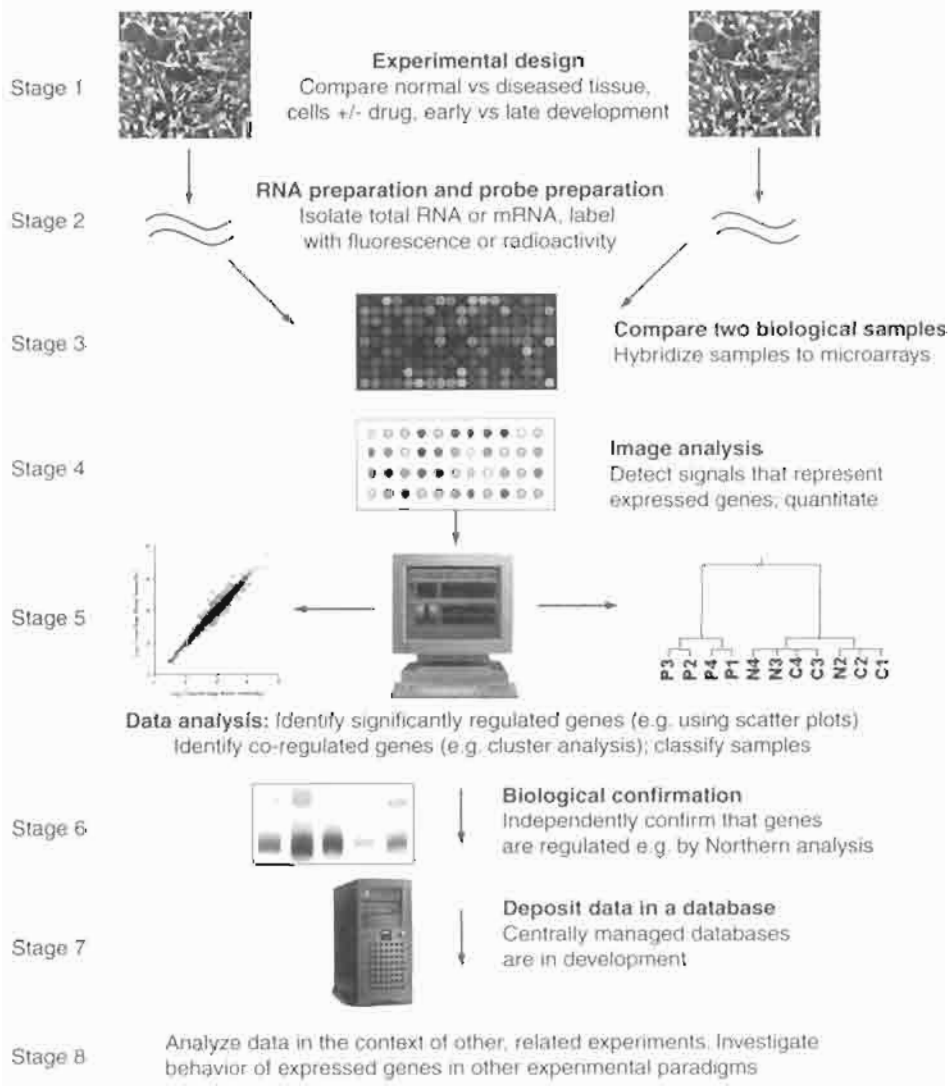


FIGURE 6.16. *Overview of the process of generating high throughput gene expression data using microarrays. In stage 1, biological samples are selected for a comparison of gene expression. In stage 2, RNA is isolated and labeled, often with fluorescent dyes. These samples are hybridized to microarrays, which are solid supports containing complementary DNA or oligonucleotides corresponding to known genes or ESTs. In stage 4, image analysis is performed to evaluate signal intensities. In stage 5, the expression data are analyzed to identify differentially regulated genes (e.g. using scatter plots; stage 5, at left) or clustering of genes and/or samples (right). Based on these findings, independent confirmation of microarray-based findings is performed (stage 6). The microarray data are deposited in a database (stage 7) so that large-scale analyses can be performed (stage 8).*

Stage 1: Experimental Design for Microarrays

In the first stage, total RNA or mRNA is isolated from samples. Notably, experiments have been described for organisms as diverse as viruses, bacteria, fungi, and humans (see Chapters 13–17). The amount of starting material that is required is typically several hundreds of milligrams (wet weight) or several flasks of cells. For many currently available microarrays, about 1–25 μg of total RNA is required. With the amplification of RNA or cDNA products, it is possible to use substantially less starting material. However, the amplified population may not faithfully represent the original RNA population.

The experimental design of a microarray experiment can be considered in three parts (Churchill, 2002). Different sources of variation are associated with each of these three areas.

We will discuss experimental design further in Chapter 7.

[1] First, the biological samples are selected for comparison, such as a cell line with or without drug treatment. If multiple biological samples are used, these are called "biological replicates."When experimental subjects are selected for treatment, it is appropriate to assign them to groups randomly.

[2] Second, RNA is extracted and labeled (typically as complementary DNA) with radioactivity or fluorescence. When two RNA extractions are obtained from a biological sample and hybridized to microarrays, these are called "technical replicates." For technologies in which one sample is hybridized to one microarray, the transcripts are labeled with radioactivity or with a fluorescent dye. Figure 6.17

For microarrays from Affymetrix, RNA is converted to cDNA and transcribed to make biotin-labeled complementary RNA (cRNA).

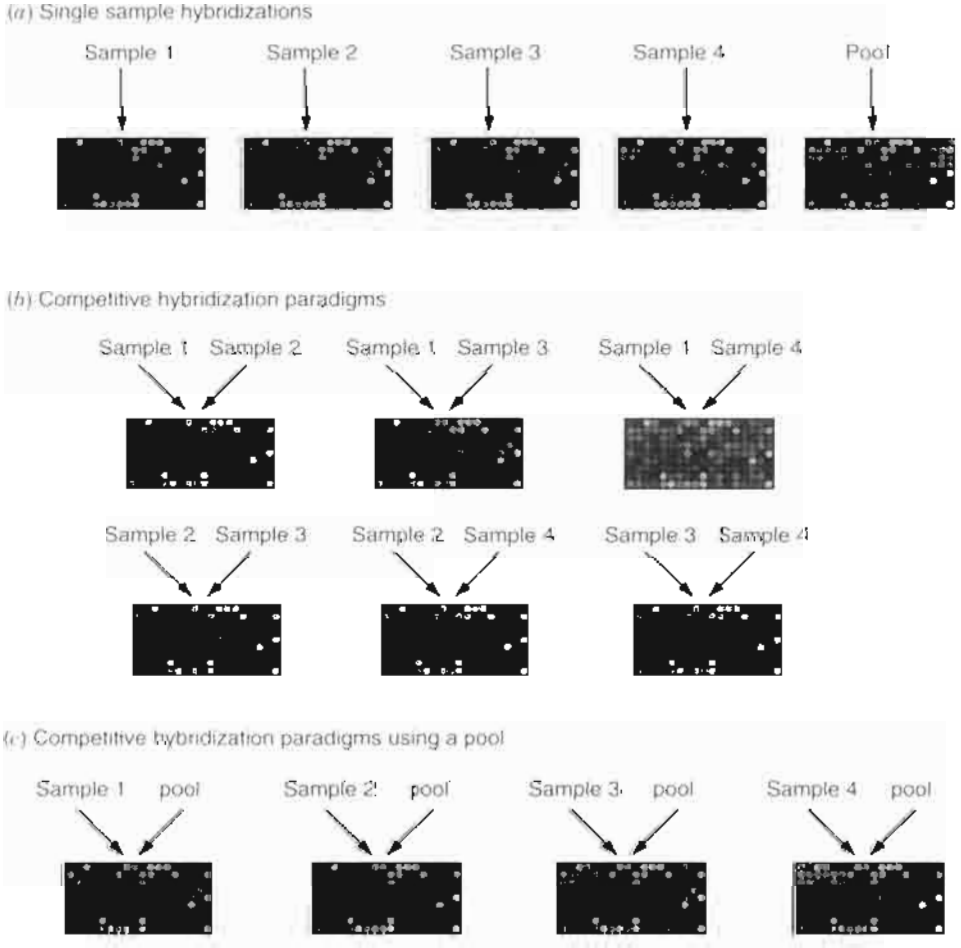


FIGURE 6.17. *Designs of microarray experiments. The four samples may represent a time course, normal versus diseased tissue, or any other paradigm. (a) Design of arrays to which only one sample is hybridized (e.g., nylon filters from Clontech that are probed with radiolabeled cDNA or chips from Affymetrix that are probed with cRNA probes that are visualized with a fluorescent dye). Each sample is hybridized to one chip. The use of a pool of all the samples is not necessary. Additional hybridizations may be performed to increase the number of replicates. (b) For technologies that use competitive hybridization, such as NEN Life Sciences arrays, samples are labeled with Cy5 (red) and Cy3 (green) dyes and competitively hybridized. A set of hybridizations may be performed to compare every combination of samples. The data do not allow intensities from one chip (e.g., sample 1 vs. 2) to be compared across chips. (c) In order to perform data analysis of every sample compared to every other sample, it is also possible to hybridize each sample to a pool consisting of a reference.*

(panel A) shows an example of an experimental design in which five samples are hybridized to arrays. Four of these samples may represent time points in an assay, different pharmacological treatments, or any other paradigm. Optionally, a reference pool consisting of multiple samples from the same treatment condition may be employed.

The experimental design differs for microarray technologies that employ competitive hybridization (Figure 6.17, panels B and C). Samples are labeled with Cy5 (a red dye) or Cy3 (a green dye). Each labeled molecule can bind to its cognate on the surface of the microarray. If the transcript is expressed at comparable levels in the two samples, the color of the spot will be intermediate (i.e. yellow). This approach produces ratios of gene expression measurements, rather than absolute values. The experimental design may involve pairwise comparisons of all the samples in order to allow comprehensive data analysis (Figure 6.17, panel B). Alternatively, each sample can be competitively hybridized with a reference pool, such as a mixture of all four samples. (Figure 6.17, panel C). Churchill (2002) and others have suggested that reference pools can represent an inefficient experimental design.

Standardization and Normalization of Microarray Data (SNOMAD) is a freely available web-based collection of algorithms for the analysis of microarray data (Colantuoni et al. 2002b). It includes a correction for spatial artifacts based on local regression analysis. The program calculates the background on an array on a local basis. See the discussion of microarray normalization in Chapter 7.

For any two-color experimental paradigm, dye swap is an important control. For two biological samples, two hybridizations are performed. Samples are labeled twice, first with Cy5 and then with Cy3, and used in independent hybridizations. The dye swap helps to eliminate artifactual variation that is attributable to the efficiency of dye labeling.

[3] A third aspect of microarray experimental design is the arrangement of array elements on a slide. Ideally, the array elements are arranged in a randomized order on the slide. In some cases, array elements are spotted in duplicate (see Figure 6.19 on page 180). Artifacts can occur based on the arrangement of elements on an array, or because a microarray surface is not washed (or dried) evenly.

Stage 2: RNA Preparation and Probe Preparation

RNA can be readily purified from cells or tissues using reagents such as TRIzol (Invitrogen). For some microarray applications, further purification of RNA to mRNA [poly(A)$^+$ RNA] is necessary. In comparing two samples (e.g., cells with or without a drug), it is essential to purify RNA under closely similar conditions. For example, for cells in culture, conditions such as days in culture and percent confluence must be controlled for.

The purity and quality of RNA should also be assessed spectrophotometrically (by measuring $a260/a280$ ratio) and by gel electrophoresis. Fluorescent dyes such as RiboGreen (Molecular Probes) can be used to quantitate yields. Purity of RNA may also be confirmed by Northern analysis or PCR. RNA preparations that are contaminated with genomic DNA, rRNA, mitochondrial DNA, carbohydrates, or other macromolecules may be responsible for impure probes that give high backgrounds or other experimental artifacts.

The RNA is labeled with radioactivity or fluorescence to permit detection.

Stage 3: Hybridization of Labeled Samples to DNA Microarrays

A microarray is a solid support such as a nylon membrane or glass slide to which DNA fragments of known sequence are immobilized. In some cases, the immobilized DNA consists of approximately 5 ng of cDNA (length 100–2000 bp)

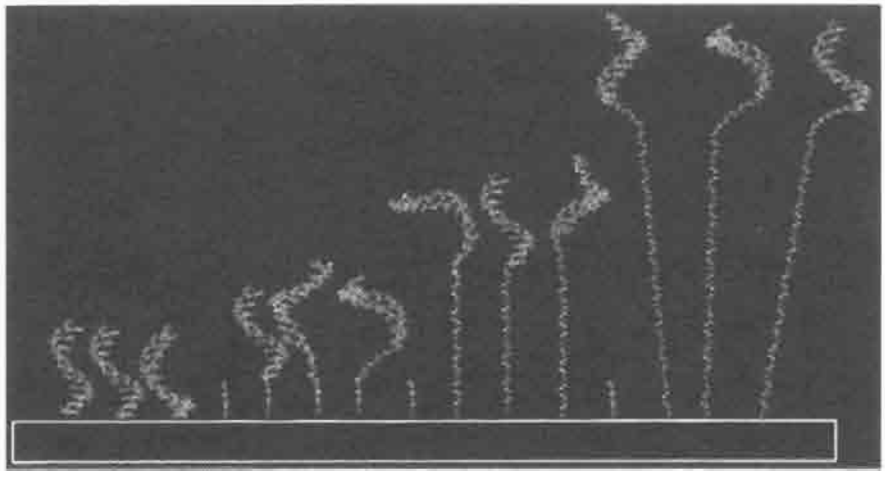


FIGURE 6.18. *The surface of a typical microarray chip contains oligonucleotides at a density of 0.1 pmol/mm^2 on a glass slide or one molecule per 39 Å^2 (from Southern et al., 1999). A typical microarray from Affymetrix contains 20 separate oligonucleotide 25-mers, each corresponding to a single gene. An additional set of oligonucleotide 25-mers incorporates a mismatch that is expected to fail to hybridize to the cRNA species from the biological sample. The extent to which an endogenous transcript has been expressed in a sample is assessed by analyzing the fluorescence signal from all 40 oligonucleotides corresponding to that expressed gene. Other microarray platforms, such as arrays from Agilent or Clontech Laboratories, employ oligonucleotides up to 80 nucleotides long. Used with permission.*

arrayed in rows and columns. In other cases, oligonucleotides rather than cDNAs are immobilized (Lipshutz et al., 1999). This has been accomplished by Affymetrix using a modified process of photolithography (Fodor et al., 1991). An example of this in Figure 6.18 shows the density of oligonucleotides on the surface of a chip. Depending on the nature of the solid support used to immobilize DNA, the microarray is often called a blot, membrane, chip, or slide. The DNA on a microarray is referred to as "target DNA." In a typical microarray experiment, the gene expression patterns from two samples are compared. RNA from each sample is labeled with fluorescence or radioactivity to generate a "probe."

Photolithography is a technique with many applications, including the microelectronics industry, in which substances are deposited on a solid support. For microarray technology, oligonucleotides are synthesized in situ on a silicon surface by combinding standard oligonucleotide synthesis protocols with photolabile nucleotides that permit thousands of specific oligonucleotides to be immobilized to a chip surface.

After RNA is converted into cDNA or cRNA labeled with fluorescence or radioactivity, the efficient labeling of probe must be confirmed. This is followed by hybridization of the probe overnight to the filter or slide and washing of the microarray. The next stage is image analysis.

It is useful to compare the features of microarray experiments using probes labeled with radioactivity versus fluorescence. Radioactive probes are thought to be more sensitive, as the signal may be enhanced by extended periods of phosphorimaging. Also, 1–25 μg of total RNA is often used for radioactive cDNA synthesis. In a typical scenario, a filter is probed with one sample, stripped, then reprobed with a second sample. In contrast, fluorescent probes are thought to provide a broader dynamic range of low- and high-expressed transcripts. Fluorescent probe synthesis sometimes requires far more RNA (e.g., 100 μg of total RNA or 1–5 μg of mRNA), although some protocols require only several micrograms of total RNA.

Many researchers refer to the DNA on a microarray as the probe and the labeled DNA derived from a biological sample as the target. Thus there are opposite definitions of probe and target, and the research community has not reached a consensus. We will generally call the labeled material derived from RNA or mRNA the "probe."

Stage 4: Image Analysis

After washing, image analysis is performed to obtain a quantitative description of the extent to which each mRNA in the sample is expressed (Duggan et al., 1999).

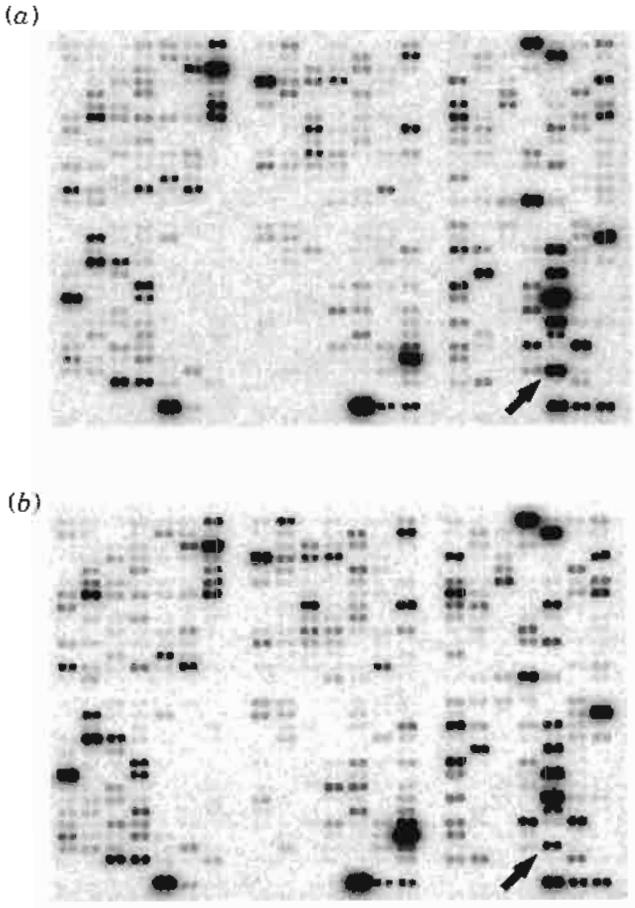


FIGURE 6.19. Example of a microarray experiment using radioactive probes. Five hundred and eighty-eight genes are represented on each array and are spotted in adjacent pairs. Dark dots represent genes expressed at high levels. The filters were hybridized, washed, and exposed to a phosphorimager screen for 6 hours. The output includes a quantitation (in pixel units) of the signals. (a) Clontech Atlas Neurobiology array probed with cDNA derived from the postmortem brain of a girl with Rett syndrome, and (b) the profile from a matched control. The arrows point at a gene (β-crystallin) that is up regulated in the disease. Note that overall the gene expression profiles appear similar in the two brains.

For experiments using radioactive probes (typically using [^{33}P] or [^{32}P] isotopes), image analysis is performed by quantitative phosphorimaging (Figs. 6.19 and 6.20, panel A). Companies that sell radioactivity-based microarrays offer software that is suitable for image analysis. This involves aligning the pixels to a grid and manually adjusting the grid to align the spots. Each spot represents the expression level of an individual gene. The intensity of a spot is presumed to correlate with the amount of mRNA in the sample. However, many artifacts are possible. The spot may not have a uniform shape. An intense signal may "bleed" to a neighboring spot, artifactually lending it added signal intensity (Fig. 6.20, panel B). Pixel intensities near background may lead to spuriously high ratios. For example, if a control value is 100 units above background and an experimental value is 200 units, this suggests that the experimental condition is up regulated twofold. However, if the pixel values are 50,100 versus 50,200, then no regulation is described.

For fluorescence-based microarrays, the array is excited by a laser and fluorescence intensities are measured (Fig. 6.20, panel C). Data for Cy5 and Cy3 channels may be sequentially obtained and used to obtain gene expression ratios.

Stage 5: Data Analysis

Analysis of microarray data is performed to identify individual genes that have been differentially regulated. It is also used to identify broad patterns of gene expression.

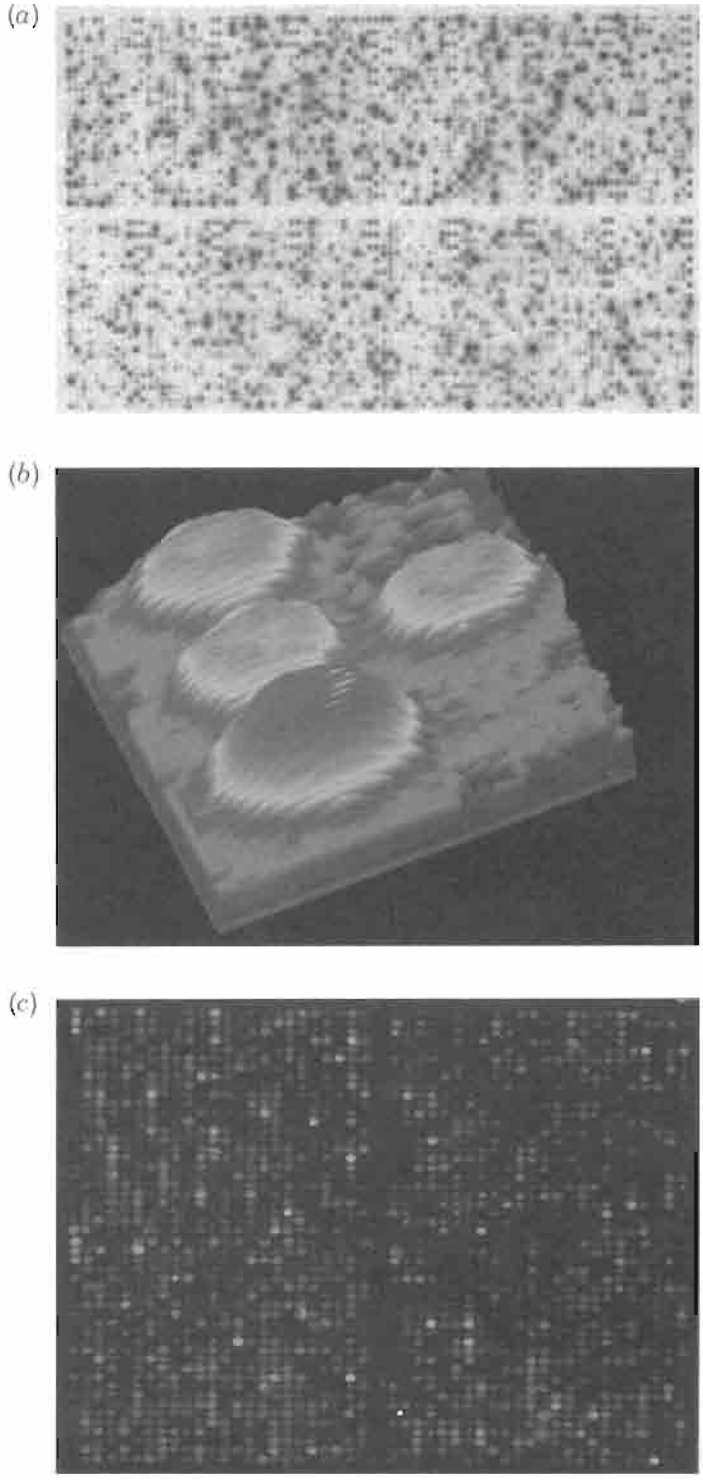


FIGURE 6.20. *GeneFilter from Research Genetics is probed with [^{32}P]cDNA derived from the hippocampus of a postmortem brain of an individual with Down syndrome. (a) There are 5000 cDNAs spotted on the array. The pattern in which genes are represented on any array must be randomized. (b) Six of the signals are visualized using NIH Image software. Image analysis software must define the properties of each signal, including the likelihood that an intense signal (lower left) will "bleed" onto a weak signal (lower right). (c) A microarray from (a) NEN Perkin-Elmer (MICROMAX, representing 2400 genes) was probed with the same Rett syndrome and control brain samples used in Figure 6.19. This technology employs cDNA samples that are fluorescently labeled in a competitive hybridization.*

In some experiments groups of genes are coregulated, suggesting functional relatedness. Samples (rather than genes) may be analyzed and classified into discrete groups. The analysis of microarray data will be described in Chapter 7.

In an effort to standardize microarray data analysis, Alvis Brazma and colleagues (2001) at 17 different institutions have proposed a system for storing and

The MIAME project is described at the Microarray Gene Expression Database Group website (http://www.mged.org/).

sharing microarray data. Minimum Information About a Microarray Experiment (MIAME) provides a framework for researchers to describe information in six areas: the experimental design, the microarray design, the samples (and how they are prepared), the hybridization procedures, the image analysis, and the controls used for normalization.

Stage 6: Biological Confirmation

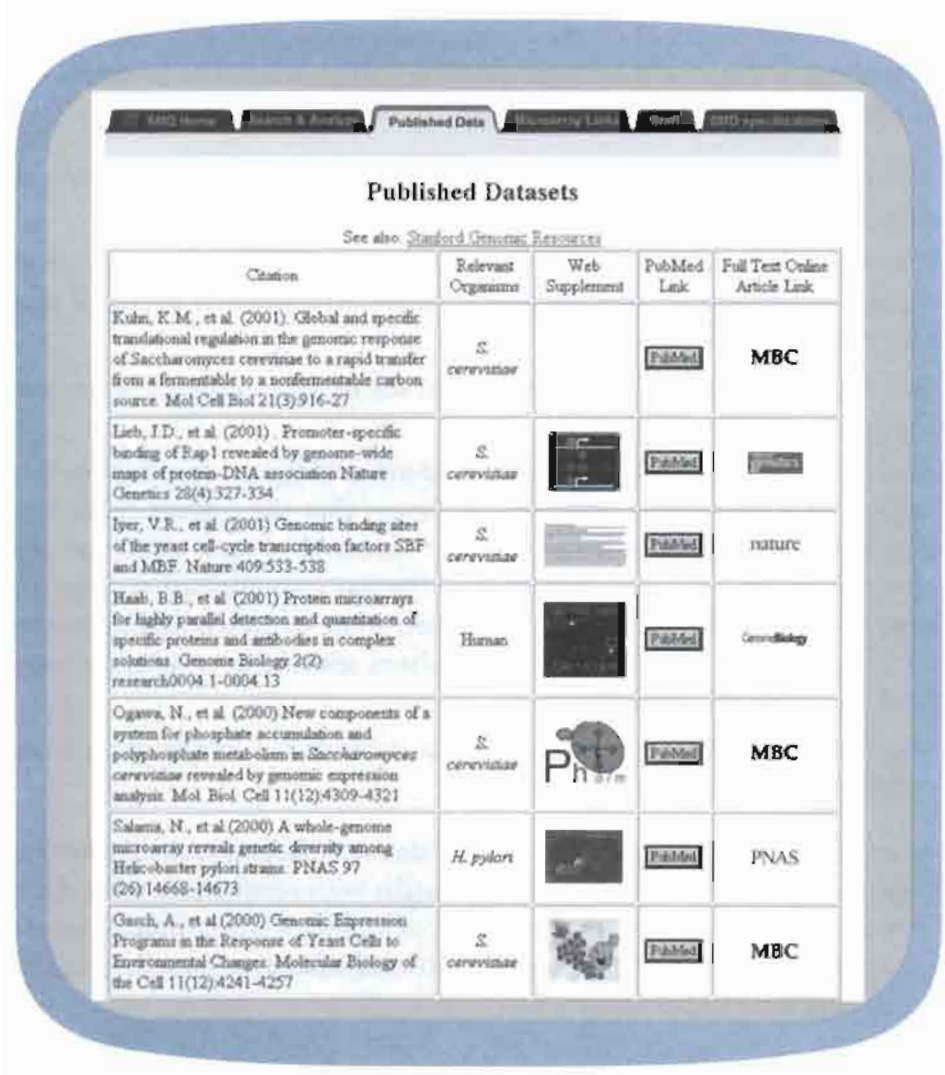
Microarray experiments result in the quantitative measurement of thousands of gene expression values. Data analysis typically reveals that dozens or hundreds of genes are significantly regulated, depending on the particular experimental paradigm and the statistical analysis approach. If one considers the top 1% of regulated genes to be "significantly" regulated, then for an array with 10,000 elements there will be 100 significantly regulated genes. Expression of these genes may be authentically regulated in cells or the genes may have been identified due to experimental artifact or random fluctuation. Thus, it is essential to independently confirm the differential regulation of at least some of the most regulated genes.

Stage 7: Microarray Databases

Many academic researchers agree that microarray data should be deposited in public repositories upon publication. Such resources are just beginning to be generated (reviewed in Gardiner-Garden and Littlejohn, 2001) (see Table 6.6 under Web Resources). Some have proposed the creation of gene expression databases that incorporate both microarray data and expression data collected using complementary technologies (e.g., Strachan et al., 1997; Aach et al., 2000). The main initial efforts have been from the European Bioinformatics Institute and from NCBI. An example of a repository is the Stanford Microarray Database, which offers links to the complete raw and processed data sets from a variety of microarray experiments (Fig. 6.21).

Stage 8: Further Analyses

Eventually, it is likely that uniform standards will be adopted for all microarray experiments. The greatest variables in these studies are likely to be the quality of the RNA isolated by each investigator and the nature of the microarray that is used to generate data. An ongoing trend in the field of bioinformatics is the unification and cross-referencing of many databases, such as has occurred for databases of molecular sequences and for databases of protein domains. In the arena of gene expression, the lack of acceptable standards may limit the extent to which an integrated view of gene expression is obtained. Nonetheless, it is likely that each gene in each organism will be indexed so that in addition to "stable" data on molecular sequence and chromosomal location, "dynamic" information on the mRNA corresponding to each gene will be cataloged. This information will include the abundance level of each transcript, the temporal and regional locations of gene expression, and other information on the behavior of gene expression in a variety of states. Some initial efforts to integrate information on gene expression are presented in Chapter 7.



Published Datasets

See also: Stanford Genomic Resources

Citation	Relevant Organisms	Web Supplement	PubMed Link	Full Text Online Article Link
Kuhn, K.M., et al. (2001). Global and specific translational regulation in the genomic response of Saccharomyces cerevisiae to a rapid transfer from a fermentable to a nonfermentable carbon source. Mol Cell Biol 21(3):916-27	S. cerevisiae		PubMed	MBC
Lieb, J.D., et al. (2001). Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nature Genetics 28(4):327-334	S. cerevisiae		PubMed	
Iyer, V.R., et al. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409:533-538	S. cerevisiae		PubMed	nature
Haab, B.B., et al. (2001) Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. Genome Biology 2(2) research0004.1-0004.13	Human		PubMed	
Ogawa, N., et al. (2000) New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis. Mol. Biol. Cell 11(12):4309-4321	S. cerevisiae		PubMed	MBC
Salama, N., et al. (2000) A whole-genome microarray reveals genetic diversity among Helicobacter pylori strains. PNAS 97 (26):14668-14673	H. pylori		PubMed	PNAS
Gasch, A., et al. (2000) Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. Molecular Biology of the Cell 11(12):4241-4257	S. cerevisiae		PubMed	MBC

FIGURE 6.21. *Stanford Microarray Database (http://www.dnachip.org/) offers resources such as access to published data sets (as shown above) and software for microarray image and data analysis.*

PERSPECTIVE

Genes in all organisms are expressed in a variety of developmental, environmental, or physiological conditions. The field of functional genomics refers to the high-throughput study of gene expression. Before the arrival of this new approach, the expression of one gene at a time was typically studied. Functional genomics may reveal the transcriptional program of entire genomes, allowing a global view of cellular function.

Complementary DNA microarrays and oligonucleotide-based microarrays were introduced in the mid-1990s and have emerged as a powerful and popular tool for the rapid, quantitative analysis of gene expression in a variety of biological systems. The use of microarrays is likely to increase in the near future as the number of genes represented on the arrays increases, the number of organisms represented on arrays increases, and the experimental applications of microarray technology expand.

PITFALLS

For studies of gene expression with techniques such as EST analysis, SAGE, or microarrays, there are many basic concerns. The mRNA molecules are not directly measured; rather, they are converted to cDNA, and that cDNA is analyzed by sequence

analysis or by visualization of fluorescent or radioactive tags. It is important to assess whether the amount of substance that is actually measured corresponds to the amount of mRNA in the biological sample.

- When RNA (or mRNA) is isolated, is it representative of the entire population of mRNA molecules in the cell?
- If two conditions are being compared, was the RNA isolated under exactly the same conditions? Any variations in the experimental protocol may lead to artifactual differences.
- Has degradation of the RNA occurred in any of the samples? For microarrays, there are additional concerns.
- Most researchers cannot confirm the identity of what is immobilized on the surface of a microarray. How do you know that the correct clone has been deposited? How do you know that the amounts of each cDNA or oligonucleotide on the array are comparable? Several unpublished studies indicate that 10–30% of the genes represented on some commercial microarrays are incorrect; that is, the wrong clone has been spotted or multiple clones have been inadvertantly spotted.
- What is the reliability of each spot that is measured on the surface of a microarray?

One response to these assorted concerns about microarrays is that with appropriate experimental design one may obtain results with confidence. After data analysis (Chapter 7) results in the identification of significantly regulated genes, it is essential to perform independent biochemical assays (such as Northern blotting or RT-PCR) to confirm the findings.

Web Resources

TABLE 6-6 Repositories for Microarray Data

Repository	Comment	URL
AMAD	From Stanford and the University of California at Berkeley and at San Francisco	http://www.microarrays.org/software.html
ArrayExpress	From Alvis Brazma and colleagues at the EBI	http://www.ebi.ac.uk/arrayexpress/
ChipDB	From the Whitehead Institute	http://young39.wi.mit.edu/chipdb_public/
ExpressDB	At Harvard; relational database containing yeast RNA expression data	http://arep.med.harvard.edu/ExpressDB/
Gene Director	From Biodiscovery	http://www.biodiscovery.com
GeNet	From Silicon Genetics	http://www.sigenetics.com
GeneX	From NCGR	http://genex.ncgr.org/
GEO	Gene Expression Omnibus from NCBI	http://www.ncbi.nlm.nih.gov/geo/
GXD	From the Jackson Laboratory	http://www.informatics.jax.org/
MAdb	National Cancer Institute	http://madb.nci.nih.gov
MaxdSQL	University of Manchester	http://www.bioinf.man.ac.uk/microarray/maxd
RAD	University of Pennsylvania	http://www.cbil.upenn.edu/rad2/servlet
Stanford Microarray Database	Stanford University	http://www.dnachip.org/

DISCUSSION QUESTIONS

[6-1] If you have a human cell line and you want to measure gene expression changes induced by a drug treatment, what are some of the advantages and disadvantages of using a subtraction approach versus SAGE versus microarrays? How are your answers different if you want to study gene expression in a less well characterized organism such as a parasite?

[6-2] When you use a microarray, how can you assess what has been deposited on the surface of the array? How do you know the DNA is of the length and composition that the manufacturer of the array specifies?

PROBLEMS

[6-1] Perform digital differential display:

- Go to UniGene (▸http://www.ncbi.nlm.nih.gov/UniGene/).
- Go to *Homo sapiens*.
- Click library differential display.
- Click some brain libraries, then "Accept changes."
- Choose a second pool of libraries to compare.

[6-2] Perform digital SAGE:

- Go to ▸http://www.ncbi.nlm.nih.gov/ and click Serial Analysis of Gene Expression.
- Click "Analyze by library."
- Compare two SAGE library collections (e.g., brain and ovary).
- Next, go to Entrez or LocusLink and select a DNA sequence (any will do).
- Copy the DNA to the clipboard, and return to the SAGE page.
- Click "Virtual Northern."
- Paste in your sequence, and submit.

SELF-TEST QUIZ

[6-1] The stages of RNA processing include all of the following except:
(a) Splicing
(b) Export
(c) Methylation
(d) Surveillance

[6-2] Digital differential display (DDD) is used to compare the content of expressed sequence tags (ESTs) in UniGene's cDNA libraries. ESTs are also represented on microarrays. Which statement best describes ESTs?
(a) Clusters of nonredundant sequences (approximately 500 bp in length)
(b) Stretches of DNA sequence that are repeated many times throughout the genome
(c) Sequences corresponding to expressed genes that are obtained by sequencing complementary DNAs
(d) A "tag" (i.e., a fragment of DNA) derived from complementary DNA (cDNA) that corresponds to a transcript that has not been identified

[6-3] UniGene has cluster sizes from very small (e.g., 1) to very large (e.g., >10,000). What does it mean for there to be a cluster of size 1?
(a) One sequence has been identified that has a very large number of EST transcripts (e.g., over 10,000) associated with it.
(b) One sequence has been identified that corresponds to a gene that has been expressed one time.
(c) One sequence has been identified (presumably it is an EST) that matches one other known sequence (thus allowing it to be identified as a UniGene cluster).
(d) One sequence has been identified (presumably it is an EST) that is thought to correspond to a known gene, but it matches no other known sequences in UniGene (i.e., it does not align to any other ESTs).

[6-4] In comparing two cDNA or SAGE libraries, you may identify a gene that is differentially expressed in one. Fisher's exact test can be used to test the null hypothesis that:
(a) The number of sequences for any given gene in two pools being compared (e.g., heart and liver) is the same in either pool.
(b) The number of sequences for any given gene in two pools being compared (e.g., heart and liver) differs in the two pools.
(c) The number of sequences in the two pools being compared (e.g., heart and liver) is the same.
(d) The number of sequences in the two pools being compared (e.g., heart and liver) differs.

[6-5] In analyzing cDNA libraries, a pitfall is that:
(a) The libraries may be derived from different tissues.
(b) The libraries may contain thousands of sequences.
(c) The libraries may have been normalized differently.
(d) The libraries may contain many rarely expressed transcripts.

[6-6] What advantage do oligonucleotide-based microarrays have over cDNA-based arrays?
(a) Two samples can be simultaneously and competitively hybridized to the same chip.
(b) It is easier for the experimenter to verify the identity of each gene that is represented on the array.
(c) It is possible to identify expression of alternatively spliced transcripts.
(d) They are far more sensitive.

[6-7] Most microarrays consist of a solid support on which is immobilized:
(a) DNA
(b) RNA
(c) Genes
(d) Transcripts

[6-8] RNA samples are commonly converted to cDNA or cRNA for microarray studies and visualized by labeling with:
(a) Radioactivity or phosphorescence
(b) Radioactivity or fluorescence
(c) Radioactivity or RNA probes
(d) Radioactivity or DNA probes

[6-9] The purpose of the MIAME project is to provide:
(a) A unified system for the description of microarray manufacture
(b) A unified system for the description of microarray experiments from design to hybridization to image analysis
(c) A unified system for the description of microarray probe preparation including fluorescence- and radioactivity-based approaches
(d) A unified system for microarray databases including standards for data storage, analysis, and presentation

[6-10] The expression of thousands of genes can be measured using cDNA libraries, SAGE, and DNA microarrays. A unique advantage of using DNA microarrays is that:
(a) The expression levels can be described quantitatively.
(b) It is possible to measure the expression levels of thousands of genes in two particular conditions of interest.
(c) It is more practical than the other experimental approaches to compare the expression levels of thousands of genes in two particular conditions of interest.
(d) It can be used to survey the expression levels of essentially all genes in a genome.

Suggested Reading

An early report on cDNAs by Mark Adams, J. Craig Venter, and colleagues (1991) described the sequence analysis of over 600 randomly selected human brain cDNAs. Remarkably, 337 of these represented novel genes. They subsequently identified over 6000 more human genes in further screens (Adams et al., 1992, 1993). Gregory Schuler (1997) of NCBI presents an overview of ESTs.

In this chapter we described serial analysis of gene expression (SAGE). Victor Velculescu and colleagues (1995) at Johns Hopkins introduced this technology. Figure 6.5 is derived from that paper. Velculescu and colleagues (1997) further used SAGE in an insightful study of gene expression in yeast. In the February 2001 human genome issue in *Science,* Caron et al. describe a human transcriptome map based on SAGE data, showing variations in gene expression based on chromosomal location.

There are many introductions to microarray technology (Duggan et al., 1999; Hegde et al., 2000; Zhang, 1999). The procedure by which oligonucleotides can be immobilized on solid supports is described by Stephen Fodor and colleagues (1991). This approach, adopted by Affymetrix, depends upon light-directed, spatially addressable, massively parallel chemical synthesis of oligonucleotides using techniques of photolithography. This paper describes the application of this technology to immobilizing both peptides and oligonucleotides onto surfaces.

References

Aach, J., Rindone, W., and Church, G. M. Systematic management and analysis of yeast gene expression data. *Genome Res.* **10,** 431-445 (2000).

Adams, M. D., et al. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* **252,** 1651-1656 (1991).

Adams, M. D., et al. Sequence identification of 2,375 human brain genes. *Nature* **355,** 632-634 (1992).

Adams, M. D., Kerlavage, A. R., Fields, C., and Venter, J. C. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat. Genet.* **4,** 256-267 (1993).

Bishop, J. O., Morton, J. G., Rosbash, M., and Richardson, M. Three abundance classes in Hela cell messenger RNA. *Nature* **250,** 199-204 (1974).

Bonaldo, M. F., Lennon, G., and Soares, M. B. Normalization and subtraction: Two approaches to facilitate gene discovery. *Genome Res.* **6,** 791-806 (1996).

Bouton, C. M. and Pevsner, J. Dragon: Database Referencing of Array Genes Online. *Bioinformatics* **16,** 1038-1039 (2000).

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., Gaasterland, T., Glenisson, P., Holstege, F. C., Kim, I. F., Markowitz, V., Matese, J. C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., Vingron, M. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29,** 365-371 (2001).

Caron, H., et al. The human transcriptome map: Clustering of highly expressed genes in chromosomal domains. *Science* **291,** 1289-1292 (2001).

Carulli, J. P., et al. High throughput analysis of differential gene expression. *J. Cell. Biochem. Suppl.* **31,** 286–296 (1998).

Celotto, A. M., and Graveley, B. R. Alternative splicing of the *Drosophila* Dscam pre-mRNA is both temporally and spatially regulated. *Genetics* **159,** 599–608 (2001).

Churchill, G. A. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics Suppl.* **32,** 490–495.

Claverie, J. M. Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* **21,** 1821–1832 (1999).

Colantuoni C., Henry G., Zeger S., Pevsner J. SNOMAD (Standardization and NOrmalization of MicroArray Data): web-accessible gene expression data analysis. *Bioinformatics.* **18,** 1540–1541 (2002b).

Colantuoni C., Henry G., Zeger S., Pevsner J. Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. *Biotechniques.* **32,** 1316–1320 (2002a).

DeRisi, J., et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* **14,** 457–460 (1996).

Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. Expression profiling using cDNA microarrays. *Nat. Genet.* **21,** 10–14 (1999).

Fodor, S. P., et al. Light-directed, spatially addressable parallel chemical synthesis. *Science* **251,** 767–773 (1991).

Futcher, B., Latter, G. I., Monardo, P., McLaughlin, C. S., and Garrels, J. I. A sampling of the yeast proteome. *Mol. Cell. Biol.* **19,** 7357–7368 (1999).

Gardiner-Garden, M., and Littlejohn, T. G. A comparison of microarray databases. *Brief Bioinform.* **2,** 143–158 (2001).

Greenbaum, D., Jansen, R., and Gerstein, M. Analysis of mRNA expression and protein abundance data: An approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics* **18,** 585–596 (2002).

Hastie, N. D., and Bishop, J. O. The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* **9,** 761–774 (1976).

Hegde, P., et al. A concise guide to cDNA microarray analysis. *Biotechniques* **29,** 548–550, 552–554, 556 passim (2000).

Lal, A., et al. A public database for gene expression in human cancers. *Cancer Res.* **59,** 5403–5407 (1999).

Lash, A. E., et al. SAGEmap: A public gene expression resource. *Genome Res.* **10,** 1051–1060 (2000).

Lee, Y., et al. Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA). *Genome Res.* **12,** 493–502 (2002).

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. High density synthetic oligonucleotide arrays. *Nat. Genet.* **21,** 20–24 (1999).

Maniatis, T., and Reed, R. An extensive network of coupling among gene expression machines. *Nature* **416,** 499–506 (2002).

Maquat, L. E. Molecular biology. Skiing toward nonstop mRNA decay. *Science* **295,** 2221–2222 (2002).

Miller, O. L., Hamkalo, B. A., and Thomas, C. A. Visualization of bacterial genes in action. *Science* **169,** 392–395, 1970.

Milner, R. J., and Sutcliffe, J. G. Gene expression in rat brain. *Nucleic Acids Res.* **11,** 5497–5520 (1983).

Modrek, B., and Lee, C. A genomic view of alternative splicing. *Nat. Genet.* **30,** 13–19 (2002).

Quackenbush, J., et al. The TIGR Gene Indices: Analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29,** 159–164 (2001).

Sagerstrom, C. G., Sun, B. I., and Sive, H. L. Subtractive cloning: Past, present, and future. *Annu. Rev. Biochem.* **66,** 751–783 (1997).

Schmitt, A. O., et al. Exhaustive mining of EST libraries for genes differentially expressed in normal and tumour tissues. *Nucleic Acids Res.* **27,** 4251–4260 (1999).

Schmucker, D., et al. *Drosophila* Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101,** 671–684 (2000).

Schuler, G. D. Pieces of the puzzle: Expressed sequence tags and the catalog of human genes. *J. Mol. Med.* **75,** 694–698 (1997).

Southern, E., Mir, K., and Shchepinov, M. Molecular interactions on microarrays. *Nat. Genet.* **21,** 5–9 (1999).

Stekel, D. J., Git, Y., and Falciani, F. The comparison of gene expression from multiple cDNA libraries. *Genome Res.* **10,** 2055–2061 (2000).

Strachan, T., Abitbol, M., Davidson, D., and Beckmann, J. S. A new dimension for the human genome project: Towards comprehensive expression maps. *Nat. Genet.* **16,** 126–132 (1997).

van Kampen, A. H., et al. USAGE: A web-based approach towards the analysis of SAGE data. Serial Analysis of Gene Expression. *Bioinformatics* **16,** 899–905 (2000).

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. Serial analysis of gene expression. *Science* **270,** 484–487 (1995).

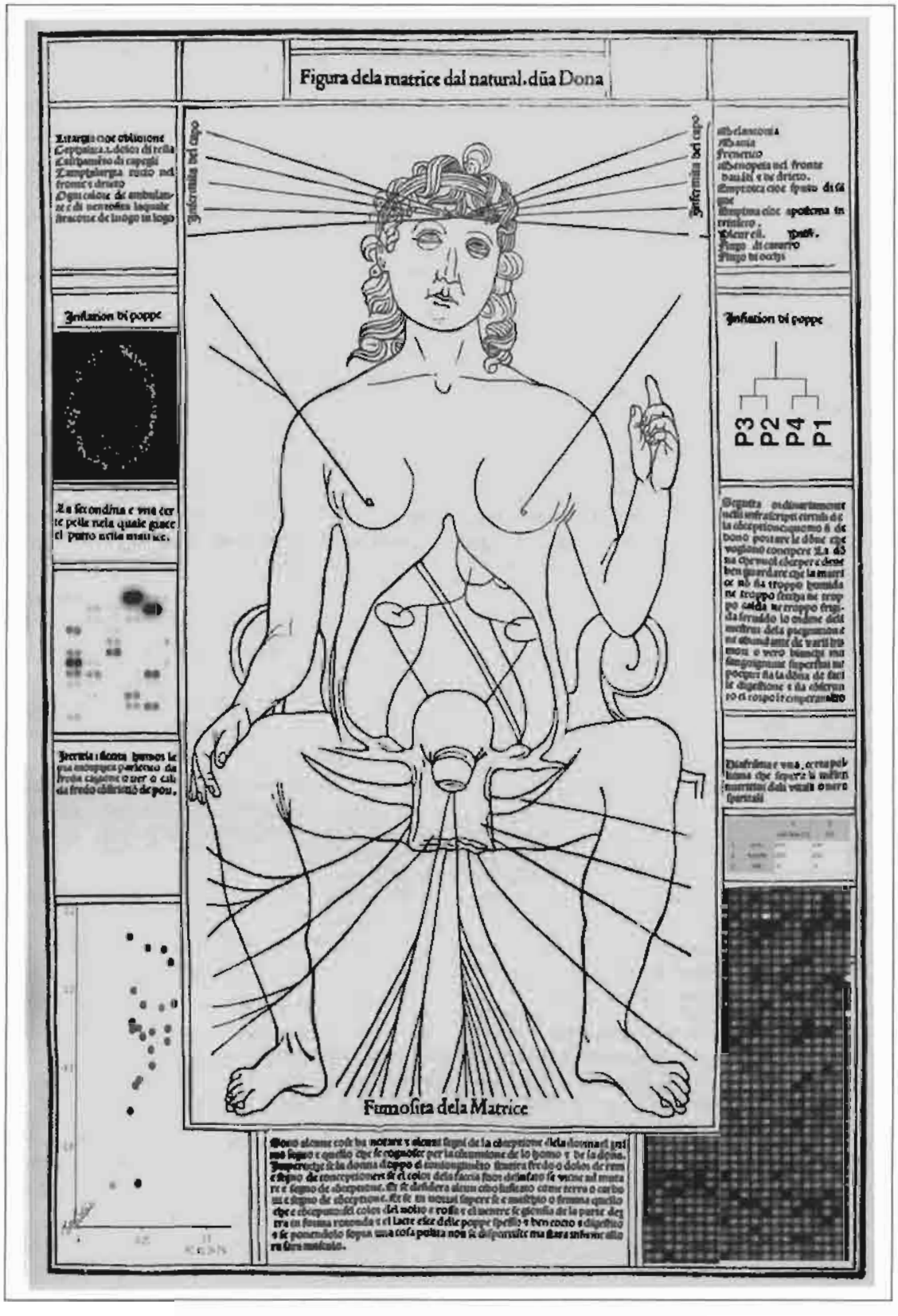
Velculescu, V. E., et al. Characterization of the yeast transcriptome. *Cell* **88,** 243–251 (1997).

Vietor, I., and Huber, L. A. In search of differentially expressed genes and proteins. *Biochim. Biophys. Acta* **1359,** 187–99 (1997).

Watson, J. B., and Margulies, J. E. Differential cDNA screening strategies to identify novel stage-specific proteins in the developing mammalian brain. *Dev. Neurosci.* **15,** 77–86 (1993).

Zhang, M. Q. Large-scale gene expression data analysis: A new challenge to computational biologists. *Genome Res.* **9,** 681–688 (1999).

Zhang, L., et al. Gene expression profiles in normal and cancer cells. *Science* **276,** 1268–1272 (1997).



Microarrays were first developed in the 1990s by the laboratories of Patrick Brown at Stanford University and Jeffrey Trent, then of the National Institutes of Health (NIH). Microarray images decorate this figure of a pregnant woman from Ketham's Fasciculo di Medicina *(Singer, 1925).*

7

Gene Expression: Microarray Data Analysis

INTRODUCTION

DNA microarray experiments have emerged as one of the most popular tools for the large-scale analyisis of gene expression. When a microarray experiment is completed and the data arrive, the first question most investigators ask is: Which genes were most dramatically up or down regulated in my experiment? This can be answered using inferential statistics, a branch of data analysis in which probabilities are assigned to the likelihood that a gene is significantly regulated:

- A spreadsheet listing all the genes represented on the array and all the expression values can be sorted to show the most differentially regulated genes.
- A scatter plot (see below) can help to quickly profile the behavior of the most regulated genes.
- A *t*-test can be used to describe the probability that a gene is regulated.

Another fundamental question that may be asked is: What signatures (or patterns or profiles) of gene expression can be found in all the gene expression values obtained in this experiment? This type of question is addressed using descriptive statistics or exploratory analysis. Clustering trees can show the relationships

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

between samples (such as normal vs. diseased cells), between genes, or both. Other tools for the analysis of gene expression include principal components analysis, multidimensional scaling, and self-organizing maps. We will consider all these tools for the analysis of array data.

Microarray experiments typically involve the measurement of the expression levels of many thousands of genes in only a few biological samples. Often, there are few technical replicates (i.e., measuring gene expression with the same starting material on independent arrays), usually because of the relatively high cost of performing microarray experiments. There are also few biological replicates (e.g., measuring gene expression from multiple cell lines, each of which has been given an experimental treatment or a control treatment) relative to the large number of genes represented on the microarray. The challenge to the biologist is to apply appropriate statistical techniques to determine which changes are relevant. There is unlikely to be a single best approach to microarray data analysis, and the tools applied to microarray data analysis are evolving rapidly.

Regardless of the microarray platform that is used, we may begin data analysis by creating a matrix of genes (along rows) and samples (arranged in columns) (Fig. 7.1, top). For microarray technologies that rely on competitive hybridization of two samples on an array, the gene expression values are ratios or relative intensities.

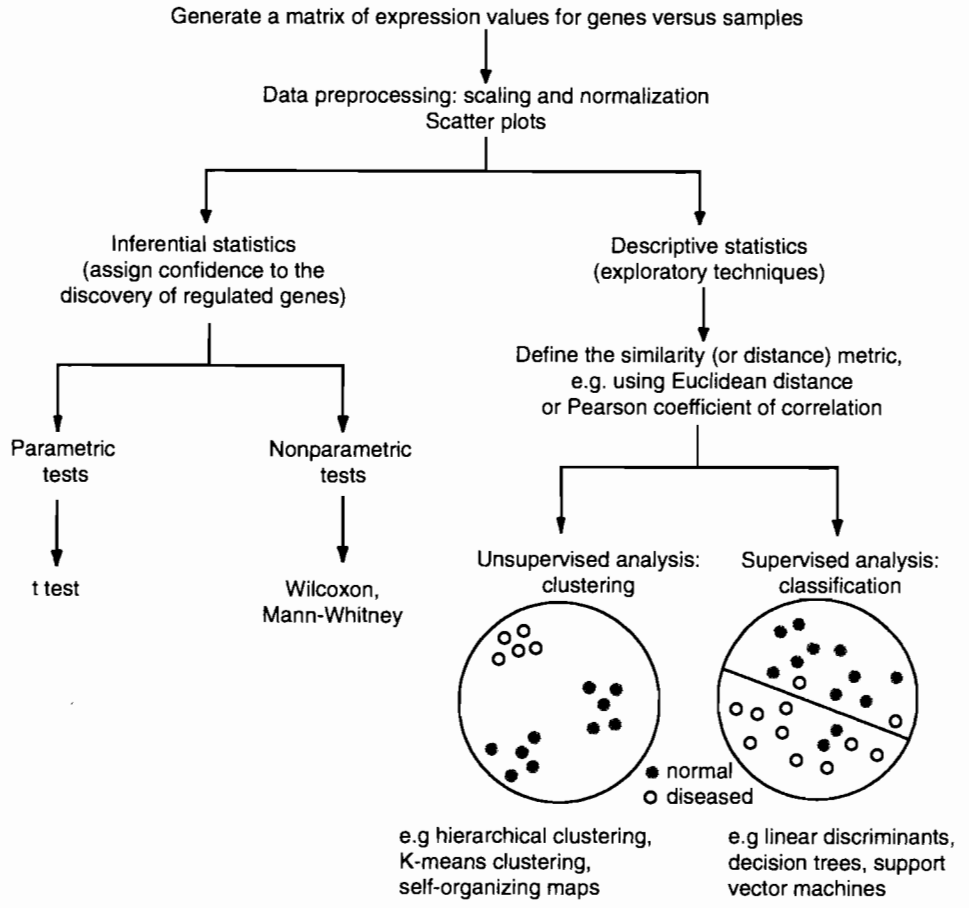


FIGURE 7.1. *Overview of microarray data analysis. A matrix of genes and samples is created and then corrected by data preprocessing (e.g., total intensity of two filters is normalized or data points with negative intensity values are eliminated). Two types of statistical analysis are performed. Inferential statistics are used to make conclusions about the data by hypothesis testing. For example, a given gene may be described as significantly up regulated, with an associated probability (p value). Alternatively, descriptive statistics may be applied in which the data set is explored. The similarities of the data points are compared with a metric such as a correlation coefficient. This pattern of gene expression may be visualized using unsupervised approaches in which patterns are sought in the representation of genes (or samples). For supervised approaches, samples (or genes) are associated with labels from a preexisting classification (such as normal vs. diseased tissue) and gene expression measurements are used to predict which unknown samples are diseased. Adapted in part from Brazma and Vilo (2000).*

Often, these are ratios of intensity values for the Cy3 (green) dye and the Cy5 (red) dye. The intensity of each signal is assumed to be directly proportional to the abundance of mRNA for each gene. In another commonly used scenario, a single sample is hybridized to a microarray. This is the case for many platforms using radioactivity-labeled cDNA or for platforms such as Affymetrix using oligonucleotides immobilized on a chip. In this case absolute values will be obtained for two (or more) experimental conditions. These absolute values can be divided for each gene to obtain ratio values.

We will describe microarray data analysis in three areas (Fig. 7.1). First, data are normalized and "preprocessed." This is essential to allow data sets from two (or more) samples to be compared to each other. Second, inferential statistics are applied. This is also called hypothesis testing, and it allows us to make statements about the likelihood that particular genes are significantly regulated. Third, exploratory statistics (also called descriptive statistics) are applied. This set of approaches includes clustering and principal-components analysis and is used to inspect the complex data set for biologically meaningful patterns. In some microarray studies classification is applied in order to diagnose physiological states (e.g., cancerous vs. control cells) based on gene expression profiles. There have been several excellent reviews of microarray data analysis (Quackenbush, 2001; Sherlock, 2001; Dopazo et al., 2001; Brazma and Vilo, 2000).

The Microarray Gene Expression Data (MGED) Society has been formed to establish standards for the analysis, annotation, exchange, and reporting of microarray data. Its website (▸http://www.mged.org) and publications by the MGED group include important information relevant to standardization of microarray experiments (Brazma, 2001; Ball et al., 2002a, b), the Miniumum Information About a Microarray Experiment (MIAME) standards (Brazma et al., 2001), the design of a microarray gene expression markup language (MAGE-ML) for standardizing the storage and exchange of microarray data (Spellman et al., 2002), and the need for public microarray data repositories (Brazma et al., 2000).

Microarray Data Analysis: Preprocessing

Gene expression changes that are identified could reflect selective, biologically relevant alterations in transcription or they could reflect variations caused by many kinds of experimental artifacts. These artifacts can include the following:

- Different labeling efficiencies of fluorescently (or radioactively) labeled nucleotides.
- Technical artifacts associated with printing tips that deposit cDNAs onto a solid support, such as uneven spotting of DNA onto the array surface.
- Variations in the performance of a fluorescence scanner (used to detect and quantitate fluorescent dyes) or phosphorimager (used for radioactivity-based arrays).
- Variations in the RNA (or mRNA) purity or quantity among the biological samples being studied. For example, there may be heterogeneity in the cell types that are dissected for studies of gene expression in a complex tissue such as the brain.
- Variations in the way the RNA is purified, labeled, and hybridized to the microarray. For example, if gene expression is measured in a cell line at two different time points and the researcher purifies different quantities of RNA, this could lead to experimental artifacts.
- Variations in the way the microarray is washed to remove nonspecific binding.
- Variations in the way the signal is measured.

The main idea of data preprocessing is to remove the systematic bias in the data as completely as possible while preserving the variation in gene expression that occurs because of biologically relevant changes in transcription (Schuchhardt et al., 2000). A basic assumption of the normalization process is that the average gene does not change in an experiment.

Terrence Speed and collaborators have developed methods for the normalization of gene expression data both within and between hybridization experiments (▸http://www.stat.Berkeley.edu/users/terry/zarray/Html/papersindex.html). Many groups have addressed basic normalization processes such as background subtraction and global mean normalization (Beissbarth et al., 2000; Hegde et al., 2000; Schuchhardt et al., 2000; Liao et al., 2000; Smid-Koopman et al., 2000; Sawitzki, 2002). Others have developed methods to identify differentially expressed genes (Claverie, 1999; Eickhoff et al., 1999; Hilsenbeck et al., 1999; Manduchi et al., 2000; Wittes and Friedman, 1999).

Global Normalization

The term "normalization" as applied to microarray data does not refer to the normal (Gaussian) distribution, but instead it refers to the process of correcting two or more data sets prior to comparing their gene expression values.

As an example of why it is necessary to normalize microarray data, the Cy3 and Cy5 dyes are incorporated into cDNA with different efficiencies. Without normalization, it would not be possible to accurately assess the relative expression of samples that are labeled with those dyes; genes that are actually expressed at comparable levels would have a ratio different than 1. Normalization is also essential to allow the comparison of gene expression across multiple microarray experiments. Thus normalization is required for both one- and two-channel microarray experiments.

As a first step, the background intensity signal is measured and subtracted from the signal for each gene (Beissbarth et al., 2000). Empty spots on the array may be used to estimate the background. This background may be constant across the surface of an array or it may vary locally. (We will discuss local background correction below.)

Most investigators apply a global normalization to raw array element intensities so that the average ratio for gene expression is 1. The main assumption of microarray data normalization is that the average gene does not change in its expression level in the biological samples being tested, and so it is reasonable for the average ratio to equal 1. The procedure for global normalization can be applied to two-channel data sets (e.g., Cy3- and Cy5-labeled samples) or one-channel data sets (e.g., Affymetrix chip data). Two-channel data are treated as two individual one-channel data sets such that each element signal intensity is divided by a correction factor specific to the channel from which it was derived. For the two or more data sets being normalized, the intensity for all the gene expression measurements in one channel (Cy5) are multiplied by a constant factor so that the total red and green intensity measurements are equal. As an example, if the mean expression value for samples in the green channel is 10,000 arbitrary units and the mean value for samples in the red channel is 5000, then the expression value for each gene in the red channel would be multiplied by 2. If the data are not log transformed, the mean ratio is then 1. (In our example, $10{,}000/10{,}000 = 1$.) If the data are log transformed (see below), the mean ratio is zero.

Other approaches to global normalization are possible. Some investigators normalize all expression values to a set of "housekeeping genes" that are represented on the array. These genes might include β-actin and glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and up to dozens of others. Then each gene expression value in a single array experiment is divided by the mean expression value of these housekeeping genes. A major assumption of this approach is that such genes do not change in their expression values between two conditions. In some cases, this assumption fails. One way to define good candidates for housekeeping genes is to analyze gene expression across a broad range of tissues and conditions. In one project, researchers measured the expression of about 7000 genes in 19 human tissues and deposited the values in a public repository, the Human Gene Expression (HuGE) Index database (Haverty et al., 2002). This database includes a list of 451 housekeeping genes that are commonly expressed across all these tissues.

You can access HuGE at ▸http://www.HugeIndex.org.

Scatter Plots

One of the most common visualization methods for microarray data is the scatter plot. This shows the comparison of gene expression values for two samples. Most data points typically fall on a 45° line, but genes that are up or down regulated fall off the line. The scatter plot rapidly displays which genes are most dramatically and differentially regulated in the experiment. A variety of software packages graph scatter plots (Fig. 7.2).

We will illustrate scatter plots in more detail using a data set from Chu et al. (1998). This group studied the developmental program of gene expression during sporulation in the budding yeast *Saccharomyces cerevisiae*. The data are easily downloadable into a spreadsheet (Box 7.1 and Fig. 7.3). Using Microsoft Excel or a variety of other graphics packages, the expression data can be graphed as a scatter plot (Fig. 7.3*b*). The main feature of this scatter plot (and most such plots of microarray data) is the substantial correlation between the expression values in the two conditions being compared. Another feature is the preponderance of low-intensity values (Fig. 7.3*b*, lower left). This means that the majority of genes are expressed at only a low level, and relatively few genes (shown at the top right of the scatter plots) are expressed at a high level.

The main goal of scatter plot analysis is to identify genes that are differentially regulated between two experimental conditions. However, the scatter plot in Figure 7.3*b* reveals more overall similarities than differences between the data sets. We may be more interested in the outliers, which could represent the most

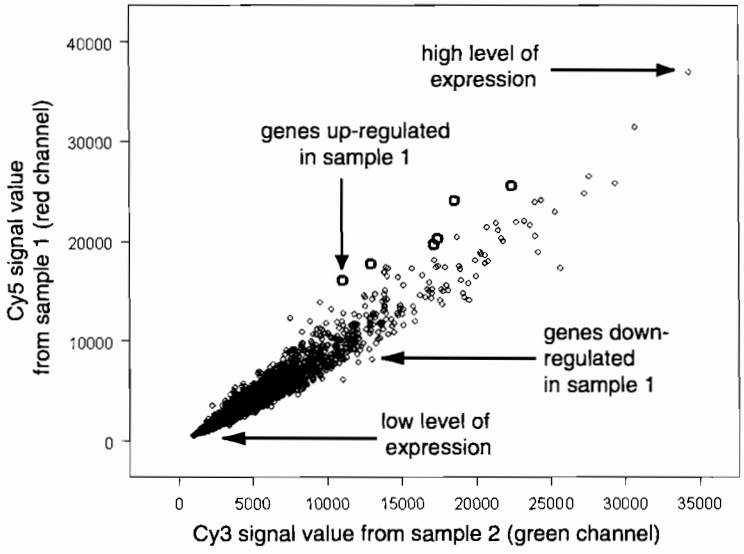


FIGURE 7.2. *Scatter plot provides one of the most basic ways of analyzing gene expression data from microarray experiments. The axes are Cy5 signal (red dye) on the y axis, here corresponding to a Rett syndrome brain sample, and Cy3 signal (green dye) on the x axis, here corresponding to a pool of control human brains. Each dot represents a gene. Genes expressed at a low level or at background are at the lower left, while genes expressed at a high level are at top right. Most data points lie along a 45° line that bisects the data. Genes with expression that are up regulated or down regulated are indicated. Note that the scale of the plot is in linear rather than logarithmic units, but data are easily converted to log scales. This plot was made using Microsoft Excel.*

BOX 7-1
Analyzing Data from a Microarray Experiment

Repositories of microarray data allow access to the raw data collected during microarray experiments, including images of the array results and tables describing the raw data (Table 6.6). One place to obtain these data is through the Stanford Microarray Database:

- Go to ▸ http://www.dnachip.org
- Click on the link to published data and find a study by Chu et al. from Patrick O. Brown and colleagues at Stanford. These data can also be accessed via ▸ http://cmgm.stanford.edu/pbrown/sporulation/.
- Select "additional figures and complete data set."
- Click "SPO spreadsheet."
- When the page loads, choose "save." You have now saved the spreadsheet of gene expression data onto a file on your computer's hard drive.
- Open the program Microsoft Excel. Open the file "spospread.txt."

Column A consists of the identifiers of yeast ORFs (open reading frames). Columns B and C contain gene expression values for green (Cy3) and red (Cy5) labeled samples at t_0 (time point zero). Additional columns in the spreadsheet correspond to further time points or to other data attributes (e.g., background values).

Select columns B and D. Use the chart wizard of Microsoft Excel to generate a scatter plot. This plot is shown in Figure 7.3 and in subsequent figures. You can input this spreadsheet into dozens of other programs such as S-PLUS (Insightful), Partek (Partek), and SAM (a Microsoft Excel plug-in described in this chapter).

dramatically regulated genes. A series of data transformations can help to identify the significantly regulated genes (Figs. 7.3–7.5). We first convert the axes from linear to logarithmic scale (Fig. 7.3*c*). The data set has an approximately symmetrical distribution about a line of slope 1. There are two main reasons to perform this transformation. First, this spreads the data from the lower left corner (Fig. 7.3*b*) to a more centered distribution in which the properties of the data set are easy to analyze. Second, it is far easier to describe the fold regulation of genes (e.g., twofold up regulated or 1.5-fold down regulated) using a logarithmic scale.

Consider the following example in which ratios are measured in a microarray experiment (Table 7.1). Gene expression values are obtained at times $t = 0, 1, 2, 3$. At $t = 1$ the relative value may be unchanged, while at time point $t = 2$ the gene is up regulated twofold and at $t = 3$ the gene is down regulated twofold. The raw ratio values are 1.0, 2.0, and 0.5. Twofold up regulation and twofold down regulation have the same magnitude of change, but in an opposite direction. In raw ratio space, the difference between 1 and 2 is +1.0, while the difference between 1 and 0.5 (i.e., time points 1 and 3) is −0.5. In log space (e.g., log base 2 space), the data points are conveniently symmetric about zero.

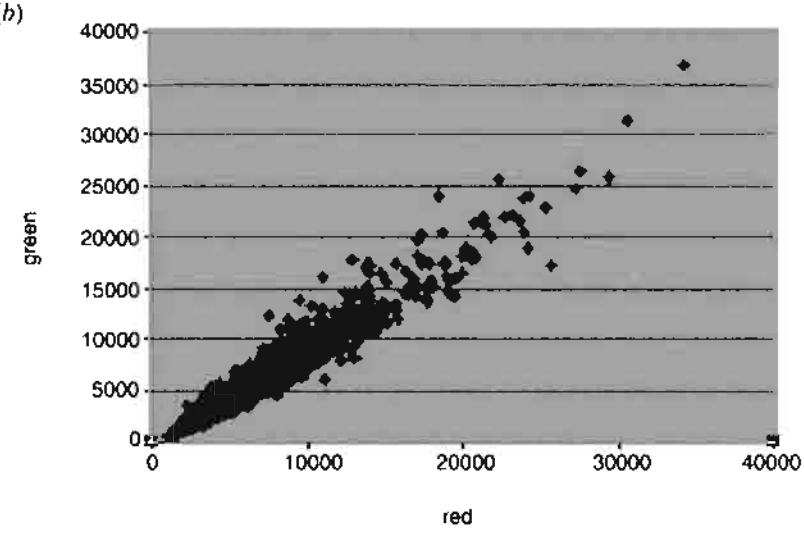
Local versus Global Normalization

Much artifactual variance present in gene expression data is not constant across the range of element signal intensities. This variation can be addressed by local

(*a*)

	A	B	C
1	ORF	t0 green	t0 red
2	YHR007C	8570	3643
3	YOL109W	7584	12218
4	YAL056W	1441	1043
5	YAL058W	2145	1740
6	YAL059W	1894	1578
7	YAL060W	7927	6770
8	YAL061W	5208	5664
9	YAL062W	8258	9527
10	YAR002W	2574	1858
11	YAR003W	2181	1897
12	YAR007C	2183	1553
13	YAR008W	1702	964
14	YAR009C	4948	4079
15	YAR010C	10550	9306
16	YAL001C	1530	1018
17	YAL002W	2802	1681

(*b*)



(*c*)

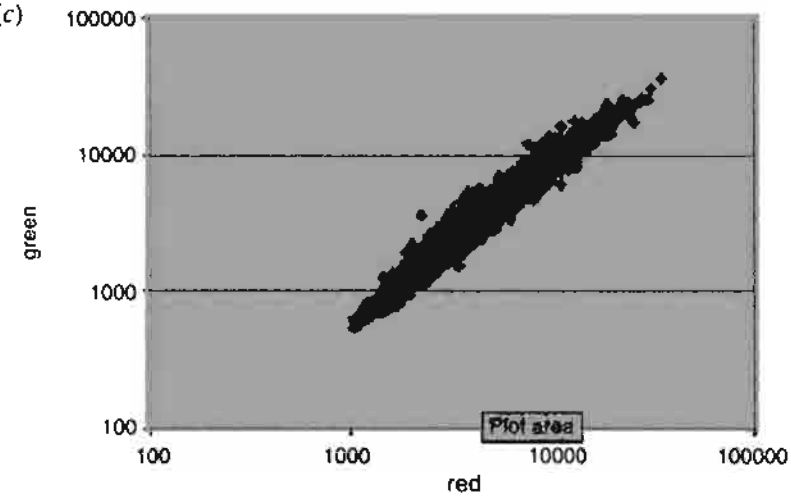


FIGURE 7.3. *Scatter plot of microarray data using Microsoft Excel. A data set from Chu et al. (1998) was downloaded from the Stanford Microarray Database (http://www.dnachip.org). (a) In this spreadsheet, columns are identifiers of yeast open reading frames and data for a time course for changes in gene expression in* S. cerevisiae *during sporulation. The rows are genes (n = 6118). Green and red refer to samples in vegetative cells versus sporulating. (b) After downloading, these data can be opened in Excel and plotted. Here, expression data for states at time point 0 are plotted. (c) Transformation of the scale to logarithmic has the effect of distributing the data points more evenly, rather than clumping most values in the lower left corner, as in (b).*

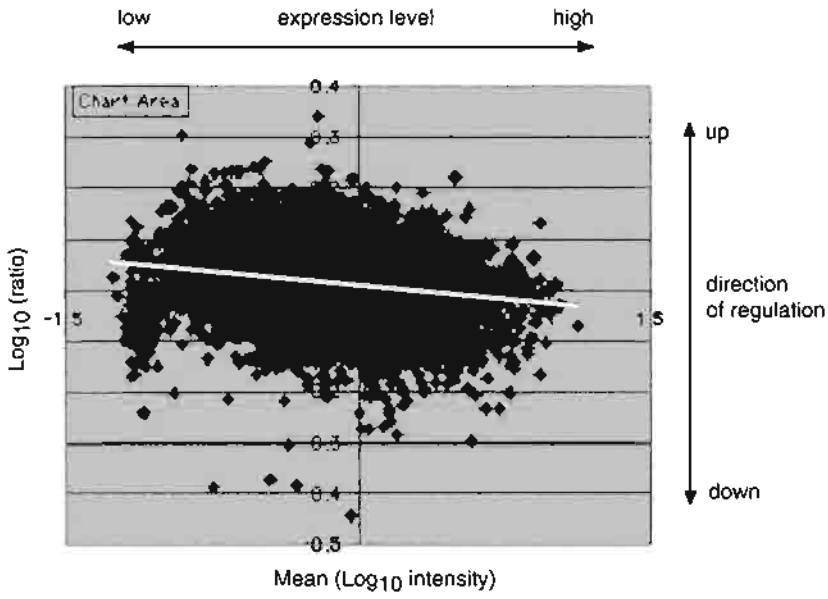


FIGURE 7.4. Transformation of scatter plot data from microarray experiments. A plot of the geometric mean intensity (x axis) versus the log of the gene expression value ratios (y axis) results in this plot of the Chu et al. (1998) data. This is related to the plot in Figure 7.3b tilted 45°. Here, the x axis reflects levels of gene expression, and the y axis reflects up or down regulation of gene expression. Some skewing of the data is revealed by performing a linear fit of the data with a local regression function, loess (white line). This skewing may be corrected (see Fig. 7.5).

normalization processes, which address bias and variance that are nonuniformly distributed across absolute signal intensity. We can continue our use of the Chu et al. (1998) data set. Perform background subtraction and global normalization of the data as described above. Next employ further sets of data transformations that are useful to allow identification of regulated genes (Fig. 7.4). Across the x axis, plot the geometric mean intensity, that is, the arithmetic mean of the $\log_{10}$ intensities of the two experiments being compared. Mean $\log_{10}$ intensity (x axis) is a measure of the average intensity of a particular element across the control and experimental conditions. This describes the magnitude of gene expression. Thus, genes expressed at a low level are now to the left, and genes expressed at a high level are to the right. The y axis is a plot of the ratio of the two intensities for each array element. In this format, it is easy to identify genes that are up regulated (greater values on the y axis) or down regulated (smaller values on the y axis). A formal definition of the x- and y-axis transformations is shown in Box 7.2.

The plot in Figure 7.4 emphasizes relative changes in element intensity (y axis) as a function of absolute level of intensity (x axis). Skewing of these data is also more readily apparent than in a standard scatter plot (Figs. 7.2 and 7.3). This skewing sometimes reflects experimental artifacts such as the contamination of one RNA source with genomic DNA or rRNA. (Such contaminating nucleic acid could bind to elements on the microarray.) Another source of artifact is the use of unequal amounts of radioactive or fluorescent probes on the microarray.

Skewing can be corrected with local normalization procedures (Fig. 7.5) (Colantuoni et al., 2002a, b). This data manipulation balances the gene expression ratios across the x axis, so that the mean expression ratio between the samples being compared is 0 (on a logarithmic scale) or 1 (on a linear scale) at all points across the range of signal intensities. This correction is accomplished by fitting a local regression curve to the data, providing a "best-fit" line. The new y axis consists of a corrected log ratio. It is created by taking each log ratio value and subtracting the local mean intensity value to obtain the residual.

The "loess" function in the *R* statistical language is used to calculate the local mean expression ratio. Global and local normalization can be accomplished online at Standardization and Normalization of Microarray Data (SNOMAD) (Colantuoni et al., 2002a, b). It is freely available at ▸http://pevsnerlab.kennedykrieger.org. Other packages such as the GeneTraffic software from Iobion (▸http://www.iobion.com) and the TIGR Microarray Data Analysis System (MIDAS); ▸http://www.tigr.org/software/tm4) also use loess corrections.

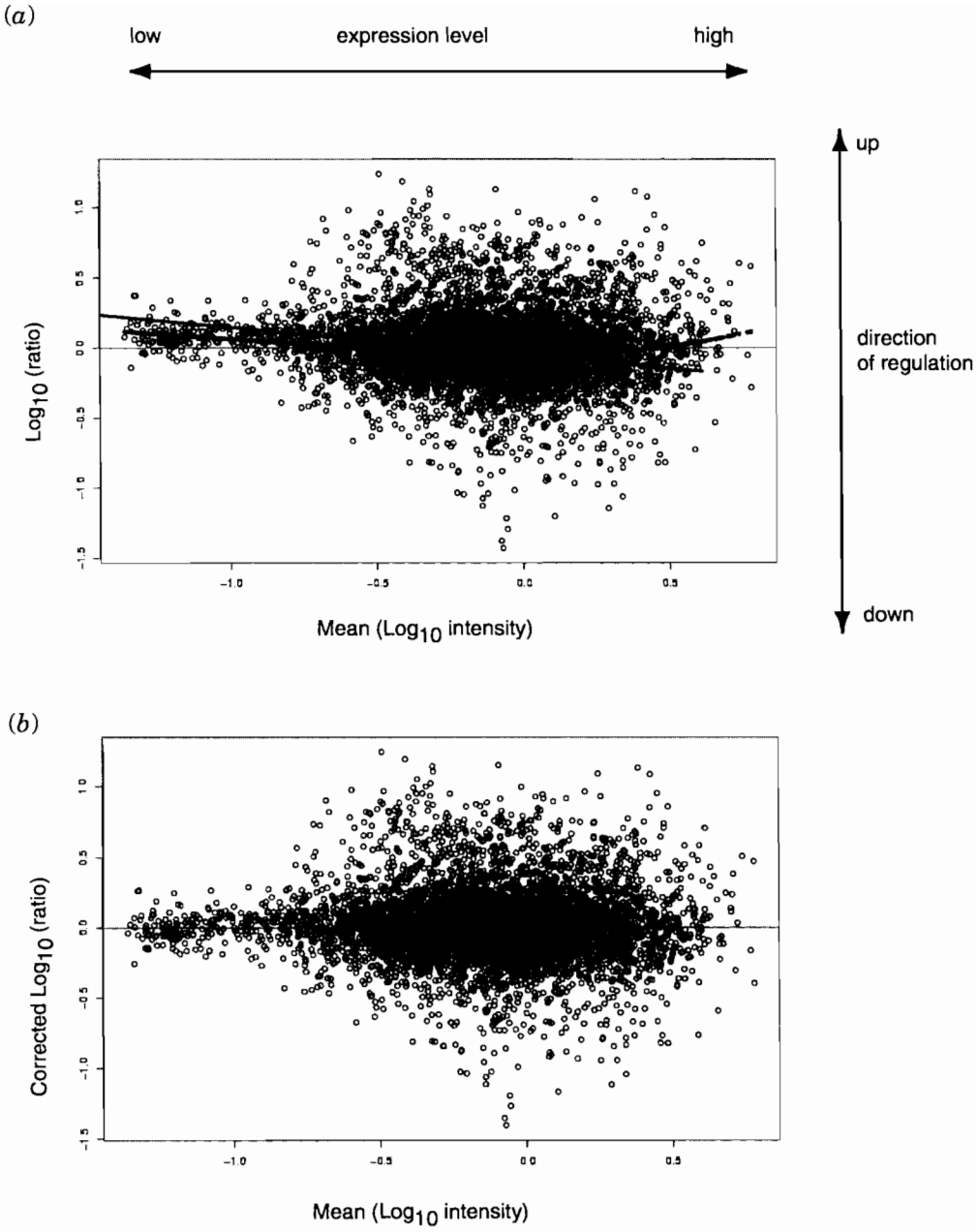


FIGURE 7.5. *Local background correction of scatter plot data from microarray experiments. (a) Plot of the log mean intensity (x axis) versus the log of the gene expression value ratios (y axis) (similar to Fig. 7.4). Skewing of the data is revealed by performing a linear fit of the data. This skewing may be corrected by calculating a local regression line through the data [red line in (a)]. (b) The distance from each data point to this best-fit line (called the residual distance) is used to correct the local background variance. These transformations were performed using SNOMAD (http://pevsnerlab.kennedykrieger.org/snomad.htm).*

TABLE 7-1 Ratios from Microarray Experiments

Log ratios of gene expression values are often easier to interpret than raw ratios.

Time (t)	Behavior of Gene	Raw Ratio Value	Log_2 Ratio Value
0	Basal level of expression	1.0	0.0
1	No change	1.0	0.0
2	Twofold up regulation	2.0	1.0
3	Twofold down regulation	0.5	−1.0

BOX 7-2
Data Transformations to Make Enhanced Scatter Plots

The scatter plots depicted in Figures 7.2 and 7.3 accentuate similarity rather than difference in expression levels across the two samples, control (CON) and experimental (EXP). In order to focus our analysis on differences, the logarithm of the geometric mean intensity (i.e., arithmetic mean of the $\log_{10}$ intensities, x axis) and the ratio of intensities (y axis) are calculated for each array element as follows (Fig. 7.4):

$$X = \log_{10}(\text{geometric mean intensity})$$

$$= \log_{10}[(\text{CON intensity} \times \text{EXP intensity})^{0.5}]$$

$$= \tfrac{1}{2}[\log_{10}(\text{CON intensity}) + \log_{10}(\text{EXP intensity})]$$

$$= \text{mean } \log_{10} \text{ intensity}$$

$$Y = \log_{10}(\text{ratio of intensities})$$

$$= \log_{10}(\text{EXP intensity/CON intensity})$$

$$= \log_{10}(\text{EXP intensity}) - \log_{10}(\text{CON intensity})$$

$$= \log_{10} \text{ ratio}$$

Mean $\log_{10}$ intensity (x axis) is a measure of the average intensity of a particular element across the control and experimental conditions. This is a measure of gene expression level. The $\log_{10}$ ratio (y axis) is a metric of each element's relative change in intensity between the experimental and control conditions.

MICROARRAY DATA ANALYSIS: INFERENTIAL STATISTICS

How can you decide which genes are significantly regulated in a microarray experiment? One approach is to calculate the expression ratio in control and experimental cases and to rank order the genes. You might apply an arbitrary cutoff such as a threshold of at least twofold up or down regulation and define those as genes of interest. One problem with a cutoff is that it is an arbitrary threshold. In some experiments, no genes (or few) will meet this criterion; in other experiments, there may be thousands of genes regulated more than twofold in either direction. Also, if the background signal level of a microarray experiment is 100 (in arbitrary units), a gene may be expressed at levels of 300 and 200 in two conditions. After background subtraction, those levels are 200 and 100, and the gene has been regulated twofold. This could have biological significance, but because the absolute values of the expression levels are so close to background, the differences could also represent noise. It is more credible that a gene that is regulated twofold with levels and 10,000 versus 5000 units is significantly regulated.

Expression ratios are important to consider, and they can quickly reveal which genes are most dramatically regulated. But these ratios cannot be converted into probability values to test the hypothesis that particular genes are significantly regulated. Many groups use expression ratios as one of several criteria to apply to

microarray data analysis. For example, Iyer et al. (1999) studied the transcriptional response of human fibroblasts to serum and selected genes with expression ratios ≥ 2.2 for subsequent cluster analysis (described below).

Another possible approach to defining which genes are significantly regulated might be to choose the 5% of genes that have the largest expression ratios. For an experiment with 10,000 gene expression values this would represent 500 genes. A problem with this approach is that it applies no measure of the extent to which a gene has a different mean expression level in the control and experimental groups. It is possible that no genes in an experiment have statistically significantly different gene expression. And yet it will always be possible to rank the genes by expression ratios and to find the group consisting of the most extreme expression ratios. Thus this approach is not useful.

The goal of inferential statistical analysis of microarray data is to test the hypothesis that some genes are differentially expressed in an experimental comparison of two or more conditions. We formulate the null hypothesis H_0 that there is no difference in signal intensity across the conditions being tested. The alternative hypothesis is that there are differences in gene expression levels. We define and calculate a test statistic which is a value that characterizes the observed gene expression data. We will accept or reject the null hypothesis based on the results of the test statistic. The probability of rejecting the null hypothesis when it is true is the significance level α, which in science is typically set at $p < 0.05$.

If you have just 100 gene expression measurements in a comparison of two groups and there is no difference in gene expression, you would expect to observe $(100)(0.05) = 5$ significantly regulated genes by chance. But assuming that these tests are statistically independent, the probability of obtaining at least one apparently significant result is $1 - 0.95^{100} = 0.994$ (see Olshen and Jain, 2002). It is for this reason that a correction needs to be applied.

The test statistic that you apply to a microarray study depends on the experimental design. Consider the basic paradigm of measuring gene expression in 10 cell lines, 5 from patients with a disease (experimental condition) and 5 from healthy individuals (control condition). You can calculate the mean and standard deviation for the expression of each gene represented on the microarray. A t-test is performed to test the null hypothesis that there is no difference in gene expression levels between the two populations. Compute the mean expression value for each gene from control (x_1) and experimental (x_2) conditions, estimate the variance (σ), and divide them:

$$t = \frac{x_1 - x_2}{\sigma} \tag{7.1}$$

The t-test essentially measures the signal-to-noise ratio in your experiment by dividing the signal (difference between the means) by the noise (variability estimated in the two groups). From the t-statistic we can calculate a p value. This allows us to either reject or accept the null hypothesis that the control and experimental conditions have equal gene expression values. An assumption of the t-test approach is that gene expression values are normally distributed.

Many commercial microarray software packages perform hypothesis testing on microarray data, and most include the option to apply conservative corrections. These packages also include a variety of data visualization tools: Partek Pro (▸http://www.partek.com), GeneSpring (▸http://www.sigenetics.com), GeneSight and Genedirector (▸http://www.biodiscovery.com), GeneTraffic (▸http://www.iobion.com/), and Spotfire (▸http://spotfire.com).

It is also necessary to apply a correction to the significance level α because of the multiple-testing problem. If you measure the expression values for 10,000 genes, you can expect to find differences in 5% of them (500 genes) purely by chance that are nominally significant at the $p < 0.05$ level. If you hypothesized that one specific gene was significantly regulated, then this α level would be appropriate. However, for 10,000 measurements it is necessary to apply some conservative correction to account for the thousands of repeated measurements you are making. Many researchers apply a conservative Bonferroni correction in which the α level for statistical significance is divided by the number of measurements taken (e.g., $p < 0.05/10{,}000$ is set as the criterion for significance).

TABLE 7-2 Test Statistics for Microarray Data

Paradigm	Parametric Test	Nonparametric Test
Compare one group to a hypothetical value	One-sample *t*-test	Wilcoxon test
Compare two unpaired groups	Unpaired *t*-test	Mann-Whitney test
Compare two paired groups	Paired *t*-test	Wilcoxon test
Compare three or more unmatched groups	One-way ANOVA	Kruskal-Wallis test
Compare three or more matched groups	Repeated-measures ANOVA	Friedman test

Source: Adapted from Motulsky (1995) and Zolman (1993).

A variety of test statistics may be applied to microarray data (e.g., Olshen and Jain, 2002). Some of these are listed in Table 7.2. These tests are all used to derive p values that help assess the likelihood that particular genes are regulated. For more than two conditions (e.g., analyzing multiple time points or measuring the effects of several drugs on gene expression), the analysis-of-variance (ANOVA) method is appropriate rather than a t-test. The ANOVA calculates the probability that several conditions all come from the same distribution.

Tests may be parametric or nonparametric. Parametric tests are applied to data sets that are sampled from a normal (Gaussian) distribution. Common parametric tests include the t-test and ANOVA. Nonparametric tests do not make assumptions about the population distribution. They rank the outcome variable (gene expression) from low to high and analyze the ranks. Nonparametric tests, such as the Mann-Whitney and Wilcoxon tests, are less commonly applied to microarray data than parametric tests.

The test that is used depends on the experimental paradigm. Some examples of experimental designs are shown in Figure 7.6. For a between-subject design (Fig. 7.6*a*) there are two groups. Golub et al. (1999) measured gene expression in samples from patients with acute leukemias that occur in two subtypes. In this experimental design it is necessary to control for confounding factors such as differences in age, gender, or weight between individuals in the two groups. For a within-subject design (Fig. 7.6*b*) a paired t-test would be used to test for the differences in mean values between two sets of measurements on paired samples. An example of this is a study by Perou et al. (2000) measuring gene expression in surgical biopsy samples before and after drug treatment of breast tumors. Here the covariates (sometimes called "nuisance variables") such as age and gender are internally controlled.

How can we be sure that the probability value we derive from a test statistic is not just obtained by chance, that is, because of random changes in gene expression? A permutation test can be performed in which the labels associated with each sample (e.g., diseased vs. control) are randomized. The same test statistic is applied to each gene, and the p value is measured. A large set of permuted tests (e.g., 100–1000) is run, and the null hypothesis is rejected if the observed p value is smaller than any p value from the permutation test.

Significance Analysis of Microarrays (SAM)

Significance analysis of microarrays (SAM) is a method that finds significantly regulated genes in microarray experiments (Tusher et al., 2001). SAM assigns a score

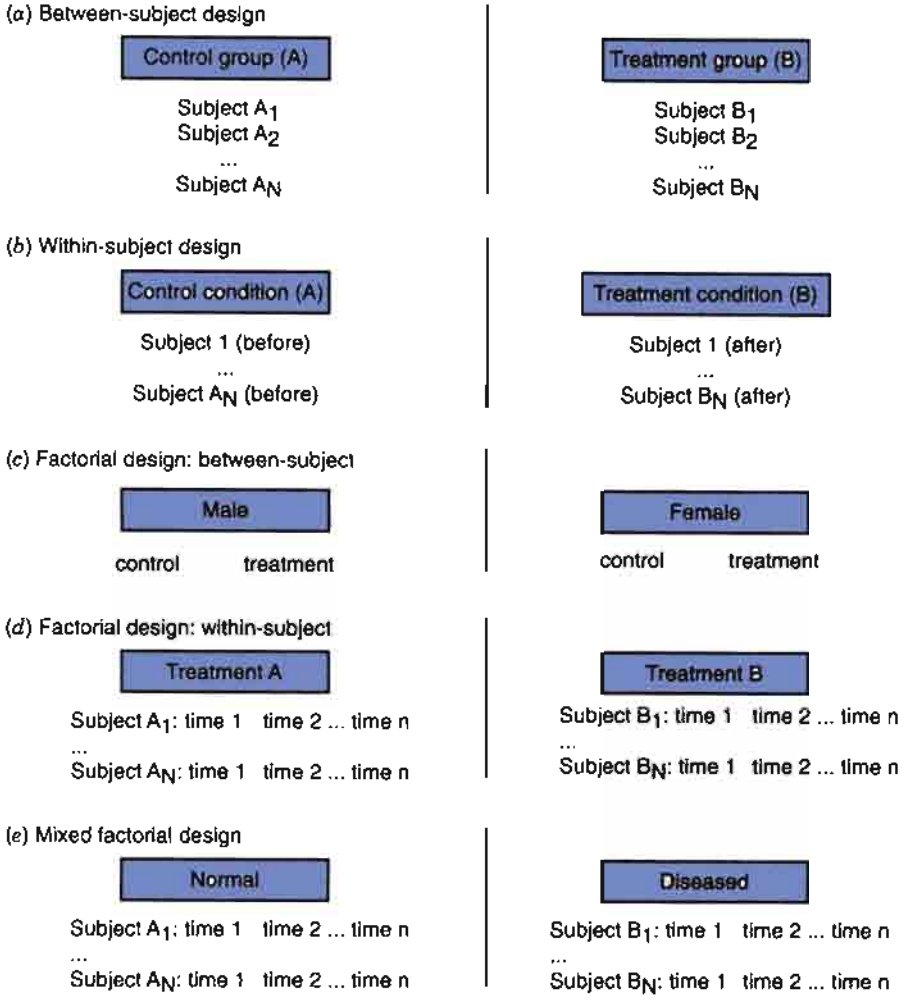


FIGURE 7.6. *Examples of experimental design for microarray experiments involving gene expression profiling. Most such microarray experiments are designed to test the hypothesis that there are significant biological gene expression differences between samples as a function of factors such as tissue type (normal vs. diseased or brain vs. liver), time, or drug treatment. (a) A between-subject design must control for confounding factors such as age, gender, or weight. (b) A within-subject design removes genetic variability and can be used to measure gene expression before then after some treatment. (c) A two-way between-subject design allows the measurement of differences between both treatment and control conditions, and another factor such as gender. (d) A within-subject factorial design might be used to study two treatments over time. (e) In a mixed factorial design there is both a between-subject design (e.g., normal vs. diseased tissue) and a within-subject design (e.g., gene expression measurements over time).*

to each gene in a microarray experiment based upon its change in gene expression relative to the standard deviation of repeated measurements.

SAM offers several useful features. The program is convenient to use as a Microsoft Excel plug-in. It accepts microarray data from experiments using a variety of experimental designs such as those outlined in Figure 7.6. Prior to operating SAM, the user must normalize and scale expression data. (This can be accomplished within Microsoft Excel or with a variety of other programs such as SNOMAD.) The SAM input data can be in a raw or log-transformed format. Each row of the data matrix contains expression values for one gene, and the columns correspond to samples. SAM uses a modified t-statistic (Equation 7.1) to test the null hypothesis (see Tusher et al., 2001).

A key feature of SAM is its ability to provide information on the false discovery rate, which is the percent of genes that are expected to be identified by chance. The user can adjust a parameter called delta to adjust the false-positive rate: for example, in a typical experiment, for every 100 genes declared significantly regulated according to the test statistic, 10 might be false positives (thus the false discovery rate would be 10%). This false-positive rate can be decreased by the user (at the cost of missing true positives) or increased (at the cost of obtaining more false negatives).

(a)

Significant Genes List

Input Parameters

Imputation Engine	10-Nearest Neighbor Imputer
Data Type	One class Response
Number of Permutations	100
Blocked Permutation?	FALSE
RNG Seed	1234567
Delta	0.52766
(Upper Cutoff, Lower Cutoff)	(0.98189, -0.88303)

Computed Quantities

Computed Exchangeability	0.78098404
50 percentile	1
False Significant Number	(5.30800, 31.73440)
False Discovery Rate (Medi	(1.47332, 7.91382)
PiHat	0.422

194 Positive Significant Genes

Row	Gene Name	Gene ID	Score(d)	Numerator(r)	Denominator(s+s0)	q-value (%)
122	GENE121	100121	2.919455217	2.667266743	0.913617968	0.1598485
499	GENE498	100498	2.87988617	2.743724628	0.952719816	0.1598485
971	GENE970	100970	2.848776207	3.479689636	1.221468231	0.1598485
508	GENE507	100507	2.835795779	2.708464332	0.954393244	0.1598485
55	GENE54	100054	2.706274015	2.591966486	0.957762027	0.1598485
982	GENE981	100981	2.624380195	2.937995936	1.119500879	0.1598485
170	GENE169	100169	2.426015119	2.261610356	0.931464693	0.1598485
501	GENE500	100500	2.410269126	2.242739373	0.930493342	0.1598485
990	GENE989	100989	2.384172951	2.550767146	1.069875046	0.1598485
524	GENE523	100523	2.322536536	2.580482728	1.111062275	0.1598485
973	GENE972	100972	2.309532469	2.396886541	1.037737539	0.1598485
795	GENE794	100794	2.294114072	2.093087158	0.912372747	0.1598485
136	GENE135	100135	2.266550033	1.989722608	0.877863969	0.1598485

(b)

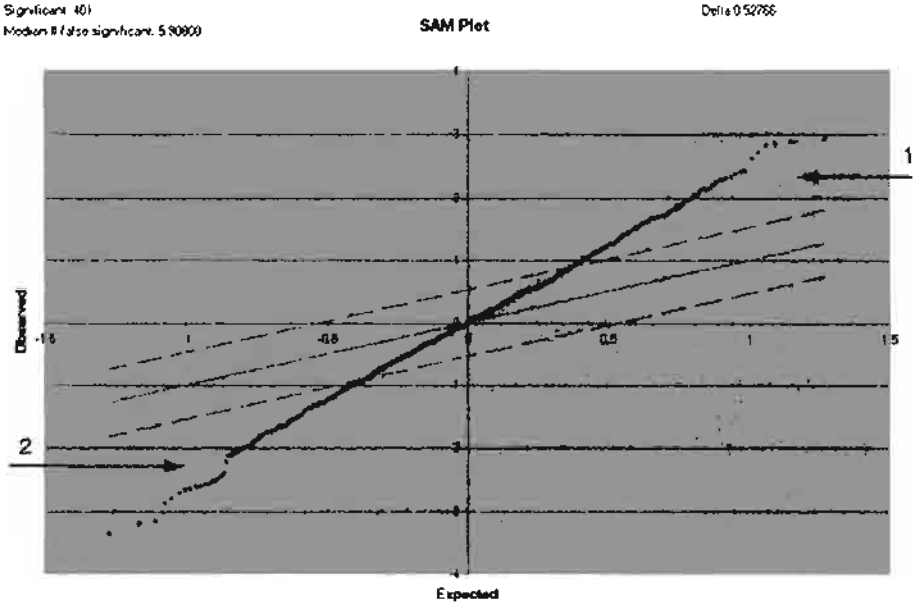


FIGURE 7.7. SAM is a Microsoft Excel plug-in that reports significantly regulated genes using a modified t-statistic. The input to SAM is a matrix of gene expression values and a response variable (e.g., control, experimental). The user selects a parameter delta to determine the cutoff for significance based on the false-positive rate. The user can also choose an appropriate fold-change measurement. (a) The output includes a list of significantly regulated genes. The score d is the t-statistic value for each gene; the numerator and denominator in the spreadsheet refer to the difference between the means of the gene expression values being compared and the estimate of the standard deviation of the numerator, respectively. The q value is the false discovery rate. (b) The output includes a plot of expected versus observed expression values; significantly up-regulated (arrow 1) and down-regulated (arrow 2) genes are color coded as well as listed in the output of (a).

You can obtain SAM at ▸http://www-stat.stanford.edu/~tibs/SAM. Note that effective permutation tests require a large number of permutations (≥100) and a reasonably large number of samples (e.g., ≥5 in each group). With too few samples, the test is not robust.

The SAM algorithm calculates a "q value," which is the lowest false discovery rate at which a gene is described as significantly regulated.

An example of a SAM output is shown in Figure 7.7. The genes are ranked according to the test statistic and plotted to show the number of observed genes versus the expected number (Fig. 7.7*b*). This graph (called a q–q plot) effectively visualizes the outlier genes that are most dramatically regulated. In SAM, a permutation test is

used to assess the significance of expressed genes; the test statistic is measured 100 or more times for each gene with the sample labels (e.g., control vs. experimental conditions) randomized.

MICROARRAY DATA ANALYSIS: DESCRIPTIVE STATISTICS

One of the most fundamental features of microarray experiments is that they generate large amounts of data. There are far more measurements (gene expression values) than samples. How can we evaluate the results of an experiment in which 10,000 gene expression values are obtained in 10 cell lines? Each gene expression value can be conceptualized as a point in 10,000-dimensional space. The human brain is not equipped to visualize highly dimensional space, and so we need to apply mathematical techniques that reduce the dimensionality of the data.

Excellent reviews of exploratory data analysis have been written by Raychaudhuri et al. (2001) and Quackenbush (2001).

Mathematicians refer to the problems associated with the study of very large numbers of variables as the "curse of dimensionality."In highly dimensional space, the distances between any two points are very large and approximately equal. Descriptive statistics are useful to explore such data. These mathematical approaches typically do not yield statistically significant results because they are not used for hypothesis testing. Rather, they are used to explore the data set and to try to find biologically meaningful patterns. A clustering tree, for example, can show how genes (or samples) form groups. Particular genes can subsequently be used for hypothesis testing.

An overview of the main descriptive techniques that are available for the visualization of microarray data is shown in Figure 7.1. Each of these approaches involves the reduction of highly dimensional data to allow conclusions to be reached about the behavior of genes and/or samples in either individual microarray experiments or multiple experiments. In each case, we begin with a matrix of genes (typically arranged in rows) and samples (typically arranged in columns). Appropriate global and/or local normalizations are applied to the data. Then some metric is defined to describe the similarity (or alternatively to describe the distance) between all the data points.

The two most commonly used metrics used to define the distance between gene expression data points are Euclidean distance (Box 7.3) and the Pearson coefficient of correlation (Box 7.4). Many software packages that perform microarray data

BOX 7-3
Euclidean Distance

Euclidean distance is defined as the distance d_{12} between two points in three-dimensional space (with coordinates x_1, x_2, x_3 and y_1, y_2, y_3) as follows:

$$d_{12} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2} \tag{7.2}$$

Euclidean distance thus is the square root of the sum of the squared differences between two features. For n-dimensional expression data, the Euclidean distance is given by

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{7.3}$$

BOX 7-4
Pearson Correlation Coefficient r

Perhaps the most common metric used to define similarity between gene expression data points is based upon the Pearson correlation coefficient r. It is used by tree-building programs such as Cluster (described below). For any two series of numbers $X = \{X_1, X_2, \ldots, X_N\}$ and $Y = \{Y_1, Y_2, \ldots, Y_N\}$,

$$r = \frac{\sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma_x} \cdot \frac{(Y_i - \bar{Y})}{\sigma_y} \right]}{(N-1)} \tag{7.4}$$

where $\bar{X}$ is the average of the values in X and σ_x is the standard deviation of these values.

For a scatter plot, r describes how well a line fits the values. The Pearson correlation coefficient always has a value between +1 (two series are identical), 0 (completely independent sets), and −1 (two sets are perfectly uncorrelated).

analysis allow you to choose between these (and other) distance measures that describe the relatedness between gene expression values.

To illustrate exploratory techniques, let us use a small portion of the Chu et al. (1998) data set. After background subtraction and global normalization, we obtain a matrix of 20 genes (rows) by three time points (columns) (Fig. 7.8*a*). The values in this matrix are log ratios. We can visualize these values in a three-dimensional plot (Fig. 7.8*b*). It is possible to see several outliers, such as genes 8 and 11 (among others). Inspecting the matrix, we can see that these two genes have relatively large expression ratios at the third time point (Fig. 7.8*a*, right column). For the full data set of 6119 gene expression values, a three-dimensional plot such as this one would not be practically useful to identify regulated genes. It is also difficult to use this plot to identify patterns in the data set such as the relative contributions of the three time points to the variance in the expression data.

The approaches we will describe first are unsupervised (Fig. 7.1). Here, prior assumptions about the genes and/or samples are not made, and the data are explored to identify groups with similar gene expression behaviors. We will then examine unsupervised clustering approaches in which you prespecify the number of clusters.

Hierarchical Cluster Analysis of Microarray Data

A single microarray experiment comparing two biological samples can be analyzed using a scatter plot, and the significance of regulated genes can be assessed using a t-test. Multiple experiments can be analyzed by with many scatter plots. Additionally, they may be analyzed by clustering of genes and/or of samples. Clustering is a commonly used tool to find patterns of gene expression in microarray experiments. Genes may be clustered in trees, or samples, or both.

Clustering is the representation of distance measurements between objects. Clusters are commonly represented in scatter plots or in dendrograms, such as those used for phylogenetic analysis (Chapter 11) or for microarray data. The main goal of clustering is to use similarity (or distance) measurements between objects to represent them. Data points within a cluster are more similar, and those in separate clusters are less similar. It is common to use a distance matrix for clustering based upon Euclidean distances.

There are several kinds of clustering techniques. The most common form for microarray analysis is hierarchical clustering, in which a sequence of nested

(*a*)

		1	2	3
		log2 t0	log2 t0.5	log2 t2
1		-0.40	-0.91	-1.60
2		-0.99	-0.07	-0.83
3		-0.22	-0.49	-0.28
4		-0.31	-0.01	-0.09
5		-0.48	1.31	0.36
6		-0.38	0.35	0.60
7		-0.41	-0.49	-0.54
8		-0.46	-2.72	-3.18
9		-0.15	0.06	0.13
10		0.12	-0.67	-0.77
11		-0.03	-1.87	-2.58
12		0.31	0.02	-1.64
13		-0.06	-0.22	0.17
14		-0.03	-0.23	0.02
15		-0.12	0.11	-0.01
16		-0.21	-0.66	-0.30
17		-0.40	1.66	1.13
18		-0.58	0.25	0.72
19		-0.77	-0.05	1.11
		-0.28	0.43	-0.57

(*b*)

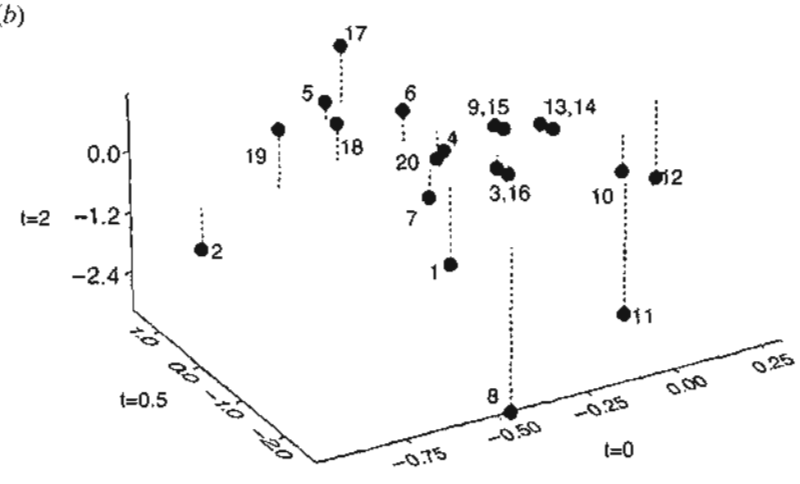


FIGURE 7.8. Matrix of gene expression values for exploratory data analysis. The set of genes in a time course experiment by Chu et al. (1998) was downloaded into Microsoft Excel (Box 7.1), the background was subtracted, the data were globally normalized, and the $\log_2$ ratios were taken for three time points (time $t = 0, 0.5, 2$). The first 20 of these gene expression ratios are displayed in (a). This data set will be used to illustrate several exploratory techniques. The data were visualized on a three-dimensional plot in (b) using the S-PLUS spreadsheet program (Insightful). The three axes correspond to the three time points. It would not be practical to generate such a plot for a large number of data points (e.g., thousands). However, in this small data set it is possible to see genes that may be differentially regulated, such as genes 2 and 8.

partitions is identified resulting in a dendrogram (tree). (We will describe a non-hierarchical clustering technique, *k*-means clustering, below.) Hierarchical clustering can be performed using agglomerative or divisive approaches (Fig. 7.9). In each case, the result is a tree that depicts the relationships between the objects. In divisive clustering, the algorithm begins at step 1 with all the data in one cluster ($k = 1$). In each subsequent step a cluster is split off, until there are n clusters. In agglomerative clustering, all the objects start apart. Thus there are n clusters at step 0; each object forms a separate cluster. In each subsequent step two clusters are merged, until only one cluster is left.

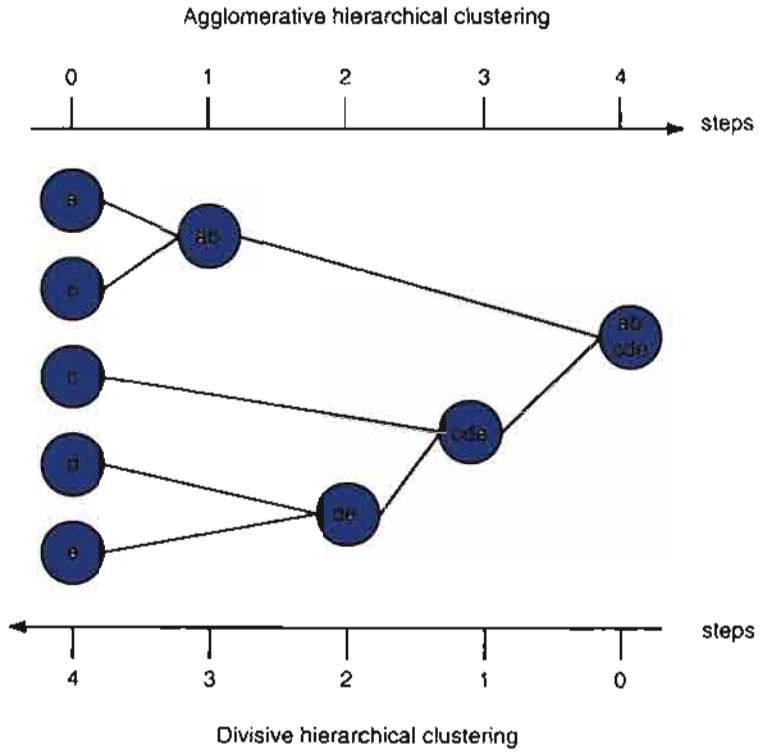


FIGURE 7.9. There are two main kinds of hierarchical clustering: agglomerative and divisive. In agglomerative clustering, the data points (genes or samples, represented as the letters a–e) are considered individually (step 0). The two most related data points are joined (circle ab, step 1). The relationship between all the data points is defined by a metric such as Euclidean distance. The next two closest data points are identified (step 2, de). This process continues (steps 3, 4) until all data points have been combined (agglomerated). The path taken to achieve this structure defines a clustering tree. Divisive hierarchical clustering involves the same process in reverse. The data points are considered as a combined group (step 0, abcde). The most dissimilar object is removed from the cluster. This process is continued until all the objects have been separated. Again, a tree is defined. In practice, agglomerative and divisive clustering strategies often result in similar trees. Adapted from Kaufman and Rousseeuw (1990). Used with permission.

The distance between clusters is most commonly defined using the average distance between all the points in one cluster and all the points in another cluster. This is called average-linkage clustering, and it is used in the unweighted pair-group method average (UPGMA). We will describe the UPGMA procedure in Chapter 11, when we define it in the context of phylogenetic trees. Other clustering algorithms are described in Quackenbush (2001).

Agglomerative and divisive clustering techniques generally produce similar results. We can use our small data set of 20 genes and three time points (Fig. 7.8) and produce two clustering trees (Fig. 7.10). For each tree, the y axis (height) represents dissimilarity. Genes 8 and 11, which we identified as possible outliers, have branches with large vertical heights. On clustering trees such as these the genes (or samples) are represented across the x axis so as to be evenly spaced, and the significance of their position depends on the cluster to which they belong. Note that while the overall topologies are similar, several of the genes are given distinctly different placements on the tree in agglomerative versus divisive clustering (Fig. 7.10, arrows). In general, different exploratory techniques may give subtle or dramatic differences in their description of the data. One way to gain confidence in a particular tree topology is to independently replicate your experiment. Another way is to examine the clusters for biological significance. If genes 1 and 12 (which are clustered adjacently in Fig. 7.10*a* but not in Fig. 7.10*b*) were both genes encoding cytokines, you might have more confidence in the agglomerative result. Yet another approach is to apply additional tests such as principal-components analysis (below).

Cluster was developed by Michael Eisen and is available through his website, http://rana.lbl.gov/

Many programs perform cluster analysis. The data in Figs. 7.8 and 7.10 were generated using S-PLUS (Insightful). One of the most popular clustering programs for microarray data is Cluster and its associated tree visualization program, TreeView.

The input for this software (and other similar programs) is a spreadsheet of expression values for genes and samples (Fig. 7.11*a*). As described above, data can be

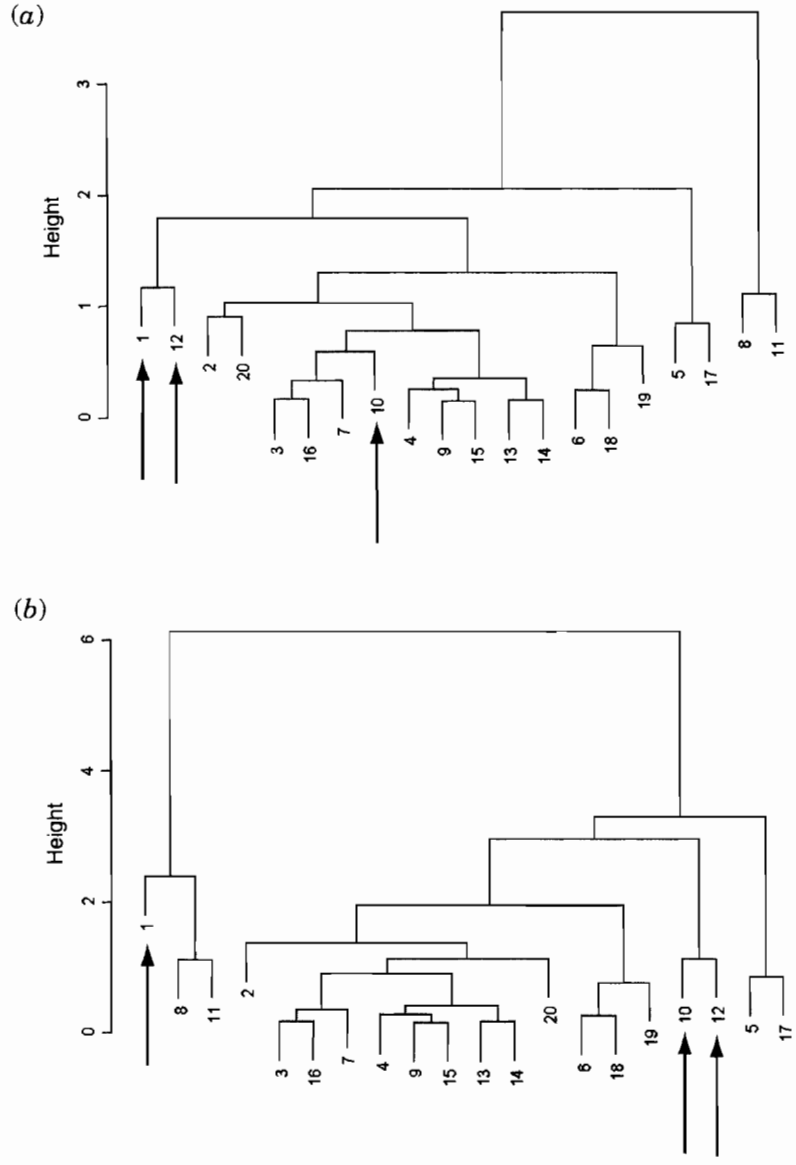


FIGURE 7.10. *(a) Agglomerative hierarchical clustering of microarray data and (b) divisive hierarchical clustering. The data set of 20 genes (Fig. 7.8a) was clustered using the S-PLUS program. The y axis reflects dissimilarity, and the genes are spaced evenly across the x axis. Note that while most of the groupings are similar between the agglomerative and divisive algorithms, there are notable differences in the placement of genes 1, 10, and 12 (arrows).*

adjusted by log transformation. Data may also be normalized to set the magnitude (sum of the squares of the values) of a row and/or column vector to 1.0. Data filtering allows genes to be removed, typically because the maximum or minimum values exceed some threshold. The distance metric used by Cluster is the Pearson correlation coefficient r, and the algorithm performs agglomerative hierarchical clustering. Examples of the output are shown in Figure 7.12. Gene expression values are color coded from bright red (most up regulated) to bright green (most down regulated). This allows one to visualize large amounts of data.

Two-way clustering of both genes and samples is used to define patterns of genes that are expressed across a variety of samples. A dramatic example is provided by Alizadeh et al. (2000), who defined subtypes of malignant lymphocytes based upon gene expression profiling (Fig. 7.13).

As another example of hierarchical clustering, we analyzed gene expression profiles of astrocytes treated with lead, sodium, or no metal. Lead is the most common environmental health threat to children in the United States, and lead exposure impairs intellectual development and causes behavioral problems. By treating

(a)

	A	B	C	D	E	F
1	YORF	0 minutes	30 minutes	1 hour	2 hours	4 hours
2	YAL001C	1	1 3	2 4	5 8	2 4
3	YAL002W	0 9	0 8	0 7	0 5	0 2
4	YAL003W	0 8	2 1	4 2	10 1	10 1
5	YAL005C	1 1	1 3	0 8		0 4
6	YAL010C	1 2	1	1 1	4 5	8 3

(b)

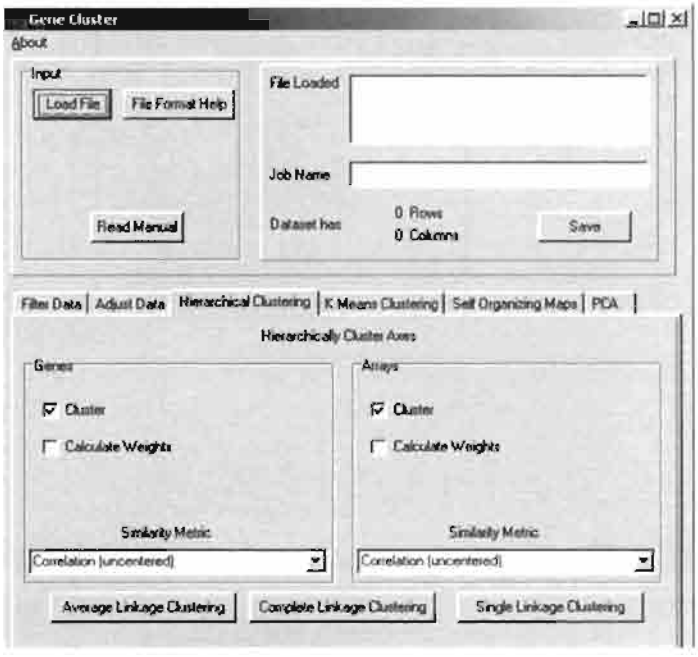


FIGURE 7.11. Two of the most popular programs for the analysis of microarray data are Cluster and Treeview, created by Michael Eisen in Patrick Brown's lab at Stanford. (a) Data are entered into Cluster as tab-delimited text files, such as those in Excel. Rows represent genes, and columns represent samples or observations. (b) The program allows a variety of analyses, including hierarchical clustering. Cluster and Treeview are available at ▸*http://rana.stanford.edu/software.*

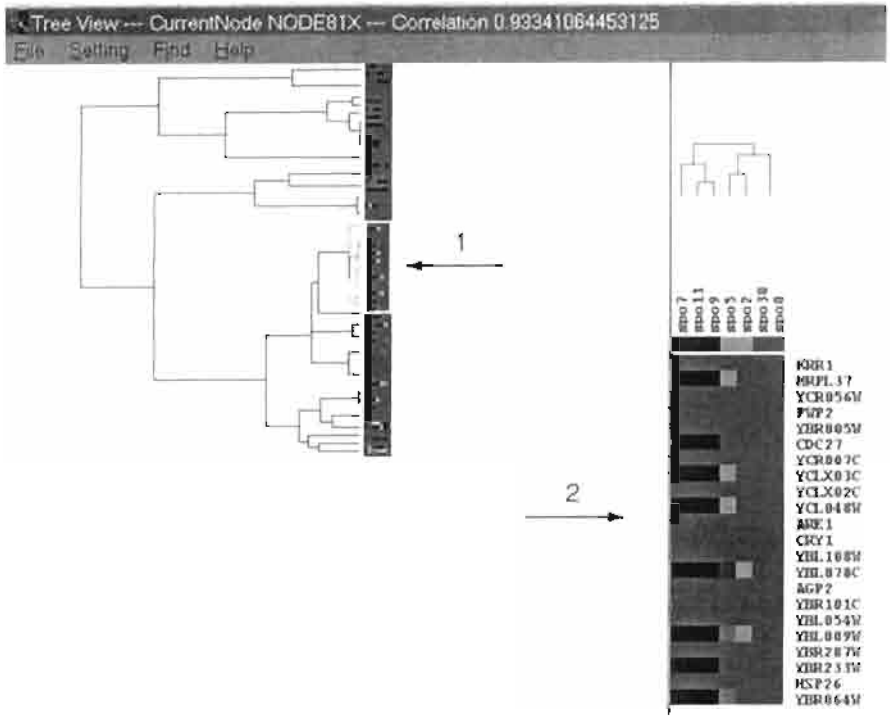


FIGURE 7.12. Examples of output from the TreeView program. The genes are color coded to aid in the visual interpretation of the output. The tree is searchable by gene name or by moving between nodes. By selecting a region of the tree (arrow 1), the corresponding genes are shown magnified at the right side (arrow 2).

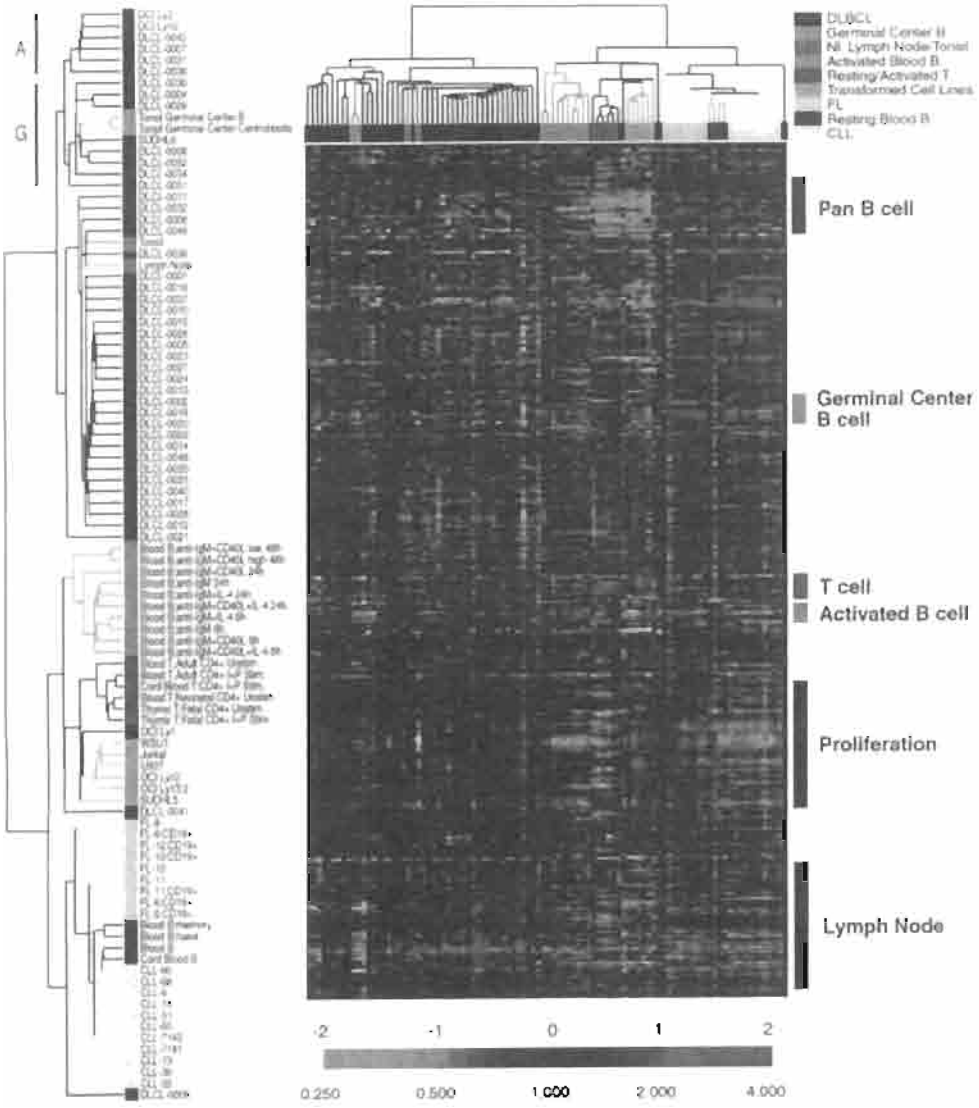


FIGURE 7.13. *Example of two-way hierarchical clustering. Alizadeh et al. (2000) made 1.8 million measurements of gene expression in 96 samples of normal and malignant lymphocytes. The cell lines are clustered in columns across the top (arrow 1), and for clarity they are also shown rotated sidewise at left (arrow 2). The genes are arranged in rows. The investigators used a custom-made microarray with 17,856 cDNA clones. This study revealed that tumors from patients with diffuse large B-cell lymphoma can be classified according to their gene expression profiles. Patients with particular tumors have varying severity of phenotype, and such heterogeneity is reflected at the molecular level. These data were generated using Cluster and Treeview. Used with permission.*

astrocytes with lead or the nontoxic metal sodium, we sought to define the molecular consequences of lead exposure to cells. One goal of the work is to define a molecular signature (or profile) for gene expression in normal versus diseased samples. Alternatively, one can cluster genes in an attempt to find coregulated genes that might share biological function. We performed cluster analysis on 11 cell lines and were able to distinguish lead-treated from control cells (Fig. 7.14) (Bouton et al., 2001). Further analysis of such trees can reveal which individual genes contribute to the ability to distinguish lead-exposed from control cells. These genes could affect cellular function and ultimately cause the behavioral deficits that occur in lead poisoning.

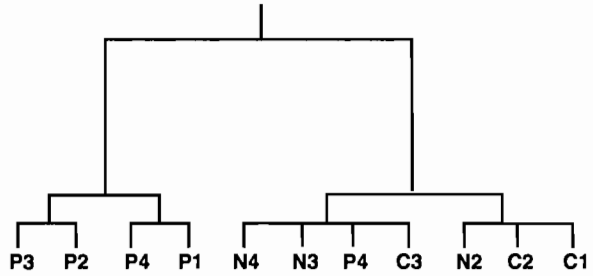


FIGURE 7.14. Example of a clustering tree. Astrocytes were treated with lead (P), which is the most common environmental health threat to children, or sodium (N), or no metal (control, C). The expression of 588 genes was measured using Clontech arrays, and the samples were clustered (Bouton et al., 2001). This shows that a signature or profile of gene expression occurred upon treatment of cells with lead. Examination of the specific genes that are differentially regulated during lead treatment could lead to an understanding of the mechanisms by which lead disrupts learning and behavior in children.

Partitioning Methods for Clustering: *k*-Means Clustering

Sometimes we know into how many clusters our data should fit. For example, we may have treatment conditions we are evaluating, or three time points. An alternative type of unsupervised clustering algorithm is a partitioning method that constructs k clusters (Tavazoie et al., 1999). Data are classified into k groups as specified by the user. Each group must contain at least one object n (e.g., gene expression value), and each object must belong to exactly one group. (In all cases, $k \leq n$.) Two different clusters cannot have any objects in common, and the k groups together constitute the full data set.

Several types of k-means clustering are performed by the S-PLUS statistical software package Partek and by Cluster/TreeView (see Web Resources, Table 7.4).

How is the value of k selected? If you perform a microarray experiment with two different kinds of diseased samples and one control sample, you might choose a value for $k = 3$. Also, k may be selected by a computer program that assesses many possible values of k. The output of k-means clustering does not include a dendrogram because the data are partitioned into groups, but without a hierarchical structure.

The k-means clustering algorithm is iterative. It begins by randomly assigning each object (e.g., gene) to a cluster. The center of each cluster is calculated (defined using a distance metric). Other cluster centers are identified by finding the data point farthest from the center(s) already chosen. Each data point is assigned to its nearest cluster. In successive iterations, the objects are reassigned to clusters in a process that minimizes the within-cluster sum of squared distances from the cluster mean. After a large number of iterations, each cluster contains genes with similar expression profiles. Tavazoie et al. (1999) described the use of k-means clustering to discover transcriptional regulatory networks in yeast.

Clustering Strategies: Self-Organizing Maps

For SOMs and other structured clustering techniques, you can estimate the number of clusters you expect (e.g., based on the number of experimental conditions) in order to decide on the initial number of clusters to use.

The self-organizing map (SOM) algorithm resembles k-means clustering. This approach to microarray data analysis has been championed by Todd Golub, Eric Lander, and colleagues from the Whitehead Institute (Tamayo et al., 1999).

Unlike k-means clustering, which is unstructured, SOMs impose a partial structure on the clusters (Fig. 7.15). The principle of SOMs is as follows. One chooses a number of "nodes" (similar to a value k) and also an initial geometry of nodes such as a 3×2 rectangular grid (indicated by solid lines in the figure connecting the nodes). Clusters are calculated in an iterative process, as in k-means clustering, with additional information from the profiles in adjacent clusters. Hypothetical trajectories of nodes as they migrate to fit data during successive iterations of SOM

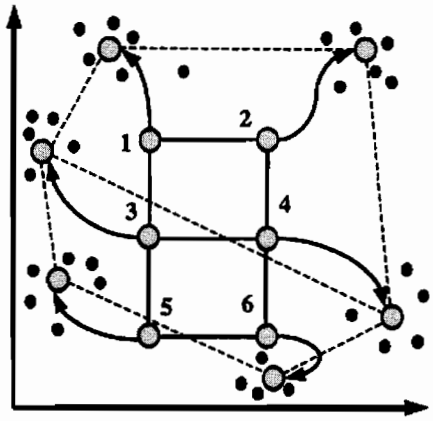


FIGURE 7.15. *Self-organizing maps allow partial structuring to be imposed on clusters. This contrasts with k-means clustering, which imposes a fixed number of clusters. A self-organizing map was used by Chu et al. (1998) as an unsupervised neural net algorithm that identified coregulated genes in sporulating yeast.*

algorithm are shown. Data points are represented by black dots, six nodes of SOM by large circles, and trajectories by arrows. The result is a clustering tree such as that shown in Fig. 7.12.

Principal Components Analysis: Visualizing Microarray Data

Principal components analysis (PCA) is an exploratory technique used to find patterns in gene expression data from microarray experiments. It is both easy to use and powerful in its ability to represent complex data sets succinctly. PCA is used to reduce the dimensionality of data sets in order to create a two- or three-dimensional plot that reflects the relatedness of the objects that it clusters—that is, the genes and/or samples in your experiment. PCA has been used to analyze expression data in yeast and mammalian systems (Landgrebe et al., 2002; Misra et al., 2002; Alter et al., 2000; Wall et al., 2001; Bouton et al., 2001).

PCA is also called singular-value decomposition (Alter et al., 2000). It is a linear projection method; this means that the data matrix you start with is "projected" or mapped onto lower dimensional space. Projection methods related to PCA include independent components analysis, factor analysis, multidimensional scaling, and correspondence analysis.

The central idea behind PCA is to transform a number of variables into a smaller number of uncorrelated variables called principal components. The variables that are operated on by PCA may be the expression of many genes (e.g., 10,000 gene expression values), or the results of gene expression across various samples, or even both gene expression values and samples. In a typical microarray experiment, the point of PCA is to detect and remove redundancies in the data (such as genes whose expression values do not change) in order to reduce the noise in the data set and to identify outliers (or clusters of outliers) that might be of interest to study.

We will consider the small matrix of 20 gene expression values in three samples (time points) from Figure 7.8*a*. We can convert this matrix into a PCA plot as shown in Figure 7.16*a*. The 20 genes are represented in the plot with numbers 1–20. Compare this plot to the clustering trees of Figure 7.10; in each case you can identify groups of genes (such as 5 and 17 or 4, 9, and 15).

The plots in Figure 7.16 were generated using S-PLUS (Insightful). Dozens of programs will perform PCA; some are listed in Table 7.4 under Web Resources.

In performing PCA, the first principal component should account for as much as the variability in the data as possible. The second principal component will account for less of the variability than the first. The mathematical operations that produce each principal-component axis require that they be orthogonal variables; this means that they are uncorrelated to each other (see below). It is typical to display PCA as a two-dimensional plot with the first principal component on the x axis and the second principal component on the y axis (as shown in Fig. 7.16*a*). However, a three-dimensional plot is also commonly used. Additional principal component axes usually account for only a very small amount of variability in the data matrix and are not displayed.

The starting point for PCA is any matrix of m observations (gene expression values) and n variables (experimental conditions). The goal is to reduce the

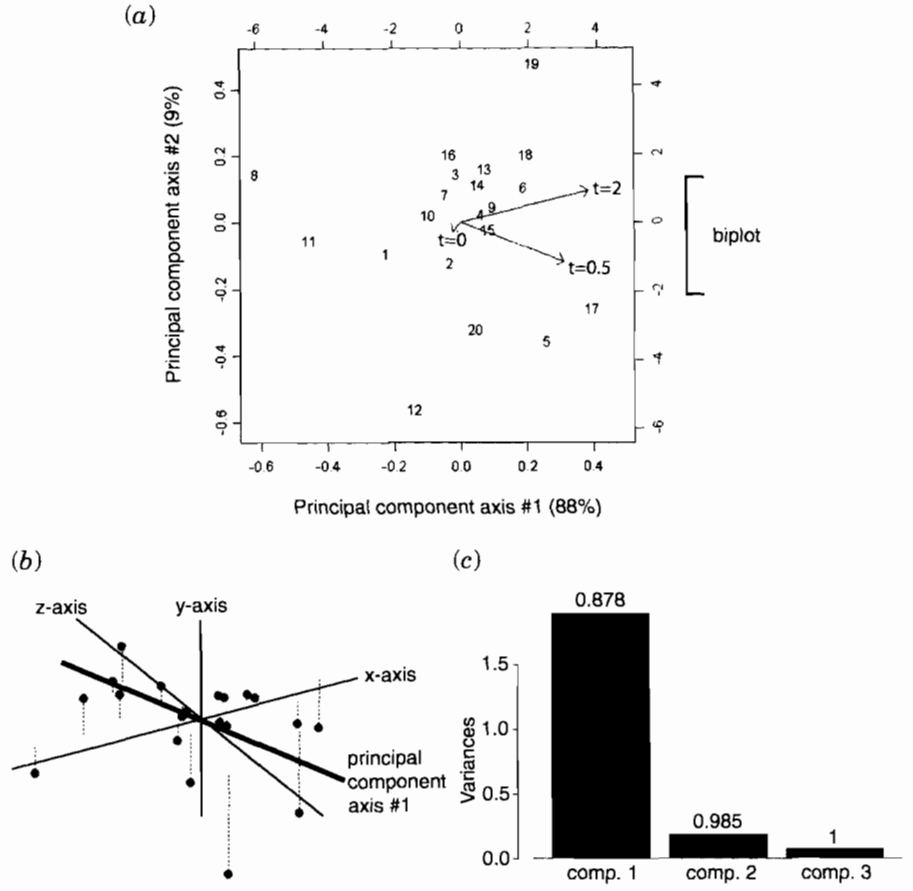


FIGURE 7.16. *PCA reduces the dimensionality of microarray data to visualize the relationship between genes or samples. A PCA plot is shown in (a). The first and second principal component axes are indicated, along with the percent of variance each accounts for. The 20 genes from the initial data set (Fig. 7.8a) are indicated on the plot; spatially close genes have similar expression profiles. A biplot is superimposed on the PCA plot and consists of three vectors. It shows the relative contribution of each of the three time points (t = 0, 0.5, 2) to the PCA plot. The first principal component axis may be thought of as the best-fit line that traverses the geometric origin of the data set in (b) accounting for most of the variability in the data. The second principal component (not shown) also passes through the origin and is orthogonal to the first component. (c) The relative importance of the three principal component axes [only two are shown in (a)] is displayed in a histogram. This shows the variances accounted for by each principal component axis as well as the cumulative percentage of variance (see numbers on top of each bar). The large percentage accounted for by the first principal component axis (88%) indicates that the importance of this axis should be weighted very heavily when inspecting the PCA plot in (a). For example, on the first principal component axis, genes 8 and 11 appear to be outliers.*

dimensionality of the data matrix by finding r new variables (where $r < n$). These r variables account for as much of the variance in the original data matrix as possible. The first step of PCA algorithms is to create a new matrix of dimensions $n \times n$. This may be a covariance matrix or a correlation matrix. (In our example in Fig. 7.16, there is a 3×3 covariance matrix.) The principal components (called eigenvectors) are selected for the biggest variances (called eigenvalues). What this means practically for our example data set is that if a gene does not vary across the three time points, it will not contribute to the formation of the principal components.

How is the first principal component axis related to our raw data? Take the three-dimensional plot of the raw data and redraw the x, y, z coordinate axes so that the origin ("centroid") is at the center of all the data points (Fig. 7.16*b*). Find the line that best fits the data; this corresponds to the first principal component axis. By rotating this axis, it becomes the x axis of Figure 7.16*a*. The second

This description of PCA is highly simplified. For a description of the vector algebra underlying PCA, see Kuruvilla et al. (2002), Misra et al. (2002), or Landgrebe et al. (2002).

principal component axis must also pass through the origin of the graph in Fig. 7.16*b*, and it must be orthogonal to the first axis. In this way, it is uncorrelated. Each axis accounts for successively less of the variability in the data. This is shown in the histogram of Figure 7.16*c*, where the first axis describes 88% of the variance. The second and third component axes account for far less (only 9 and 3%). Thus the first principal component accounts for about 10 times more of the variance than the second principal component. The third axis is not shown and accounts for the small amount of remaining variance.

A biplot is superimposed on the PCA plot of Figure 7.16*a*. This biplot is a vector drawing of the relative contributions of the three time points ($t = 0, 0.5, 2$) to this PCA analysis. There was relatively little contribution from the $t = 0$ data, and this makes sense looking at the modest amount of variation in those expression ratios (Fig. 7.8*a*).

The final product of PCA is usually a two- or three-dimensional plot consisting of points in space that correspond to either genes or samples. If we use PCA to represent samples, a close distance between any two points implies that their overall pattern of gene expression is similar. Conversely, two points that are separated in a PCA plot have different overall profiles. The initial data set is highly dimensional; for 1000 gene expression measurements the points could theoretically be described in 1000-dimensional space. PCA reduces the dimensionality of the data to just two or three dimensions. In reducing the dimensionality, the goal of PCA is to provide as much information as possible about the original data set.

The total number of dimensions in PCA can be as large as the sample size in the original data matrix, but most of the information content in PCA is found in the first two or three principal components.

Raychaudhuri and colleagues (2000) used PCA to analyze the full Chu et al. (1998) data set, consisting of 6118 gene expression measurements across seven time points. Their PCA analysis showed that just two principal components accounted for over 90% of the total variability. They further suggested that these components correspond to (1) overall induction of genes and (2) the change in induction level over time. In general, the principal component axes may or may not correspond to variables that have an obvious biological interpretation. This is because the components capture as much information in the data set as possible based strictly on the criterion of variance.

We will encounter PCA again in an entirely different context, as an application in the protein family database Pfam (Chapter 8). There, it is used to describe the relationships between proteins based upon a matrix of pairwise sequence alignments. PCA is also used to express the relationships between entire sequenced genomes in the Clusters of Orthologous Genes (COG) database (Chapter 14).

Supervised Data Analysis for Classification of Genes or Samples

The distances and similarities among gene expression values can be described using two types of analysis: supervised or unsupervised (Fig. 7.1). The unsupervised approaches we have described so far are especially useful for finding patterns in large data sets. In supervised analyses, the approach is different because the experimenter assumes some prior knowledge of the genes and/or samples in the experiment. For example, transcriptional profiling has been performed on cell lines or biopsy samples that are either normal or cancerous (e.g., Alizadeh et al., 2000; Shipp et al., 2002; Khan et al., 1998; Perou et al., 1999; West et al., 2001). (In some cases, the cancerous samples are further subdivided into those that are relatively malignant or relatively benign.) Some of these studies apply unsupervised approaches.

The goal of supervised microarray data analysis algorithms is to define a rule that can be used to assign genes or conditions into groups. In each case,

begin with gene expression values from known groups (e.g., normal vs. cancerous) and "train" the algorithm to learn a rule. Positive and negative examples are used to train the algorithm. The algorithm is then applied to unknown samples, and its accuracy as a predictor or classifier is assessed.

Some of the most commonly applied supervised data analysis algorithms are support vector machines, supervised machine learning, neural networks, and linear discriminant analysis. As an example of a supervised approach, Brown et al. (2000) used support vector machines to classify six functional classes of yeast genes: tricarboxylic acid cycle, respiration, cytoplasmic ribosomes, proteasome, histones, and helix–turn–helix proteins. They used a threefold cross-validation method: the data set is divided into thirds (sets 1, 2, and 3). Sets 1 and 2 are used to train the support vector machine; then the algorithm is tested on set 3 as the "unknowns." Next, sets 1 and 3 are used for training and set 2 is tested as the unknowns. Finally, sets 2 and 3 are used for training, and set 1 is tested. They measured the false-positive rate and found that support vector machines outperform both unsupervised clustering and alternative supervised clustering approaches.

Annotation of Microarray Data

A major task confronting the user of microarrays is to learn the biological significance of the observed gene expression patterns. Often researchers rely on manual literature searches and expert knowledge to interpret microarray results. Several software tools accept lists of accession numbers (corresponding to genes that are represented on microarrays) and provide annotation. For example, the Database Referencing of Array Genes Online (DRAGON) website allows microarray data to be annotated with data from publicly available databases such as UniGene (Fig. 7.17a) (Bouton and Pevsner, 2000; Bouton et al., 2003). DRAGON offers a suite of visualization tools allowing the user to identify gene expression changes that occur in gene or protein families (Bouton and Pevsner, 2002). The goal of annotation tools such as DRAGON is to provide insight into the biological significance of gene expression findings.

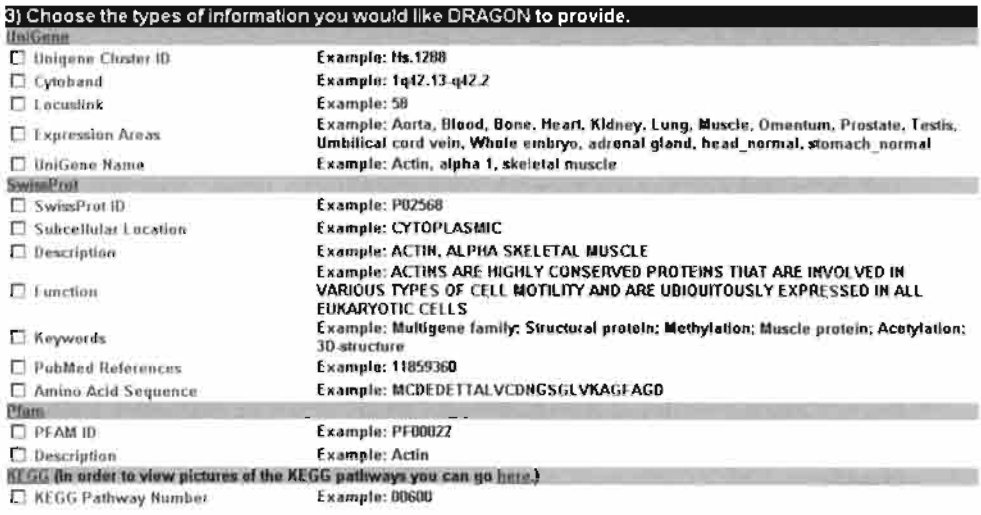
DRAGON is available at http://pevsnerlab.kennedykrieger.org. Related annotation tools include RESOURCERER at The Institute for Genomic Research (http://pga.tigr.org/tigr-scripts/magic/r1.p1), The Stanford Online Universal Resource for Clones and ESTs (SOURCE; http://source.stanford.edu) and ARROGANT (http://lethargy.swmed.edu).

PERSPECTIVE

DNA microarray technology allows the experimenter to rapidly and quantitatively measure the expression levels of thousands of genes in a biological sample. This technology emerged in the late 1990s as a tool to study diverse biological questions. Thousands to millions of data points are generated in microarray experiments. Thus microarray data analysis employs mathematical tools that have been established in other data-rich branches of science. These tools include cluster analysis, principal-components analysis, and other approaches to reduce highly dimensional data to a useful form. The main questions that microarray data analysis seeks to answer are as follows:

- For a comparison of two conditions (e.g., cell lines treated with and without a drug), which genes are dramatically and significantly regulated?
- For comparisons across multiple conditions (e.g., analyzing gene expression in 100 cell lines from normal and diseased individuals), which genes are consistently and significantly regulated?
- Is it possible to cluster data as a function of sample and/or as a function of genes?

(a)



(b)

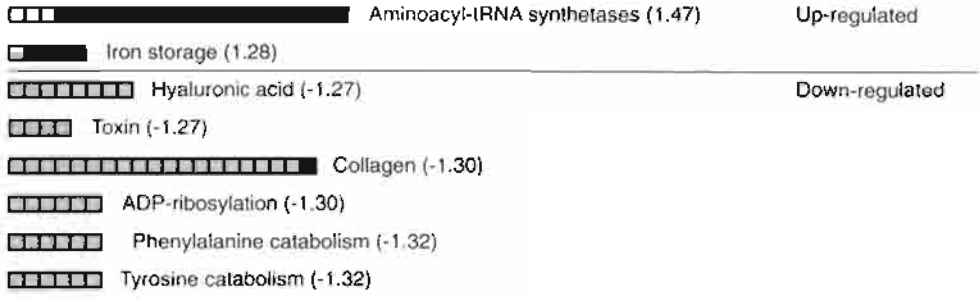


FIGURE 7.17. *Database Referencing of Array Genes Online (DRAGON) is a relational database that allows microarray data to be annotated and visualized. (a) The annotate page allows a user to input microarray data in the form of a tab-delimited text file of DNA accession numbers as well as gene expression values. The output includes data corresponding to each accession number, such as the chromosomal location of the corresponding gene or the function of the protein product. These data are obtained from databases such as UniGene, SwissProt, Pfam, and KEGG (see Chapters 8 and 10). (b) An example of DRAGON output for gene expression in cells exposed to lead. A group of genes encoding aminoacyl tRNA synthetases are up-regulated, and other families are up- or down-regulated. Each box is color-coded to indicate the degree of differential regulation, and hyperlinked to LocusLink. Adapted from Bouton et al., 2001.*

A further challenge is to translate the discoveries from microarray experiments into further insight about biological mechanisms.

Finally, while DNA microarrays have been used to measure gene expression in biological samples, they have also been used in a variety of alternative applications. Microarrays have been used as a tool to detect polymorphisms (see Chapter 17), to directly obtain the sequence of a gene, to identify regulatory DNA sequences and to identify deletions and duplications in genomic DNA (Chapter 18). Such diverse applications are likely to expand in the near future.

PITFALLS

Errors occur in a variety of stages of microarray experiments:

- Experimental design is a critical but often overlooked stage of a microarray experiment. It is essential to study an adequate number of experimental

and control samples. The appropriate number of replicates must also be employed, but there is no consensus on what this number is.

- It is difficult to relate intensity values from gene expression experiments to actual copies of mRNA transcripts in a cell. This situation arises because each step of the experiment occurs with some level of efficiency, from total RNA extraction to conversion to a probe labeled with radioactivity or fluorescence and from hybridization efficiency to variability in image analysis. Some groups have introduced universal standards for analysis of a uniform set of RNA molecules (e.g., Lucidea from Amersham Pharmacia Biotech), but these have not yet been widely adopted.
- Data analysis requires appropriate attention to global and local background correction.
- For exploratory analyses, the appropriate metric must be employed, such as Pearson's correlation coefficient, to decide whether particular genes are regulated.
- Each data analysis approach has advantages and limitations. For example, popular unsupervised methods (such as cluster analysis) sacrifice information about the classes of samples that are studied (such as cell lines derived from patients with different subtypes of cancer). Supervised methods make assumptions about classes that could be false.
- Many experimental artifacts can be revealed through careful data analysis. Skewing of scatter plots may occur because of contamination of the biological sample being studied. Cluster analysis may reveal consistent differences, not between control and experimental conditions, but between samples analyzed as a function of day or operator.

WEB RESOURCES

TABLE 7-3 Resources for Microarray Data Analysis

Resource	Comment	URL
Y. F. Leung's microarray site	List of microarray links	▸http://ihome.cuhk.edu.hk/~b400559/arraysoft.html
Stanford Microarray Database	Includes lists of software tools	▸http://www.dnachip.org/index.shtml
TIGR	Various programs	▸http://www.tigr.org/software/
Microarray & Data Analysis	Literature on array analysis	▸http://linkage.rockefeller.edu/wli/microarray/index.html
mAdb	National Cancer Institute and center for Information Technology, NIH	▸http://nciarray.nci.nih.gov/

TABLE 7-4 Microarray Data Analysis and General Statistics Packages for Analysis of Microarray Data.

Dozens of other packages are available.

Software	Company/Organization	URL
Array Suite	Affymetrix	▸http://www.affymetrix.com
Array Organizing Tool (ARROGANT)	University of Texas Southwestern Medical Center at Dallas	▸http://lethargy.swmed.edu/
Atlas Navigator	Clontech	▸http://www.clontech.com
BioArray Software Environment (BASE)	Lund University	▸http://base.thep.lu.se/

(*Continued*)

TABLE 7-4 *(Continued)*

Software	Company/Organization	URL
BioMine	Gene Network Sciences	▸http://www.gnsbiotech.com/products.shtml
Cleaver	Stanford University	▸http://classify.stanford.edu/
Cluster Analysis of Gene Expression Dynamics (CAGED)	Harvard Medical School	▸http://www.genomethods.org/caged
Cluster/Treeview	Stanford University	▸http://rana.lbl.gov/
DRAGON	Kennedy Krieger Institute	▸http://pevsnerlab.kennedykrieger.org
Expression Profiler	European Bioinformatics Institute	▸http://ep.ebi.ac.uk/
GeneCluster	Whitehead Institute	▸http://www-genome.wi.mit.edu/cancer/software/software.html
GeneCruiser	Whitehead Institute	▸http://www-genome.wi.mit.edu/cancer/software/software.html
GeneSight	BioDiscovery	▸http://www.biodiscovery.com
GeneSpring	Silicon Genetics	▸http://www.sigenetics.com
GeneTraffic	Iobion	▸http://www.iobion.com/
GeneLinker	Molecular Mining	▸http://www.molecularmining.com
J-Express	MolMine	▸http://www.molmine.com/
Partek	Partek	▸http://www.partek.com
Pathways	Research Genetics	▸http://www.resgen.com
S-PLUS	Insightful	
Signficance analysis of microarrays (SAM)	Excel plug-in	▸http://www-stat.stanford.edu/~tibs/SAM
SNOMAD	Kennedy Krieger Institute	▸http://pevsnerlab.kennedykrieger.org
Spotfire Decision Site	Spotfire	▸http://www.spotfire.com/
SPSS-Clementine	SPSS Inc.	▸http://www.spss.com/spssbi/clementine/
TIGR MEV	The Institute for Genomic Research	▸http://www.tigr.org/software/ m
X-MINER	X-Mine	▸http://www.x-mine.com/
XLSTAT	Excel plug-in	▸http://www.xlstat.com

DISCUSSION QUESTIONS

[7-1] A microarray data set can be clustered using multiple approaches, yielding different results. How can you decide which clustering results are "correct" (most biologically relevant)?

[7-2] What are the best criteria to use to decide if a gene is significantly regulated? If you apply fold change as a criterion, will there be situations in which a fold change is statistically significant but not likely to be significant in a biological sense? If you apply a conservative correction and find that no genes change significantly in their expression levels in a microarray experiment, is this a biologically plausible outcome?

PROBLEMS

[7-1] *Accessing and downloading microarray data.* Go to the Stanford Microarray Database at ▸http://www.dnachip.org:

- Click on published data. Select Chu et al. (1998), The transcriptional program of sporulation in budding yeast (*Science* 282:699–705).
- Select "Additional Figures and Complete Data Set."
- Select "Spo Spreadsheet." A text-only spreadsheet will open on the screen. Save the file to the hard drive of your computer.
- Start Microsoft Excel and open the text file that you just saved.
- Column A contains names for each of the genes on the microarrays. Columns B and D contain fluorescence intensity values for green and red labeled samples,

respectively, at time $t = 0$. Columns C and E contain background intensity values for the spots on the array.

[7-2] *Creating an Excel graph without correcting for background fluorescence.* In this portion of the analysis, we will ignore the background fluorescence measurements (columns C and E) and just work with the spot intensity values at time $t = 0$. First, raw intensities will be used; then the base 10 logarithm of each intensity will be used. See page 196 in this chapter for an explanation of the advantages of using logs.

[7-3] *Making a graph of the raw intensities:*

- Click on the B at the top of column B to highlight the entire column. Hold down the Ctrl key and click on the D at the top of column D. This will select column D while keeping B selected (but the column in between them will not be selected.)
- Click on Insert at the top of the page and choose Chart. Select XY (Scatter) and then click Next. A graph should appear, containing red intensity values on the y axis and green intensity values on the x axis. This graph is similar to the one shown in Figure 7.3*b*.
- Click Next. This box allows you to choose a title and axis labels for your graph.
- Click Next again. Here, you can decide whether to save the graph as its own sheet within the file or as an object that can be moved around on top of the spreadsheet. If you choose to save it as its own sheet, you can use the tabs at the bottom to toggle between the spreadsheet and the graph.
- Click Finish after selecting New Sheet or Object.

Note: When creating a graph, the first column that you highlight will be used as the x axis and the second column will be the y axis.

[7-4] *Making a graph of log intensities:*

- Insert a new column in the spreadsheet after column B. Do this by clicking on the C above column C, then clicking Insert at the top of the page and choosing Columns. In this column, you will calculate the base 10 logarithm of each of the intensity values in column B ("t0 green").
- Click on the second box in the new column (now column C). Move the cursor above the columns to the equal sign (=) in front of the blank line. Click once on the = sign. Then enter LOG10(B2) on the blank line. Press Enter. The number 3.552668 should appear in box C2.
- Now copy this formula to the rest of the boxes in column C. To do this, click once on box C2. Then go up to the top of the screen and click on Edit and choose Copy. Highlight all of the remaining boxes in column C and click on Edit, then Paste.
- Repeat these steps with the "t0 red" values.
- Make a scatter plot of the log columns, as shown in Figure 7.3*c*.

[7-5] *Creating an Excel graph taking background fluorescence into account.* In this portion of the analysis, incorporate the information on background fluorescence measurements (originally columns C and E, but now columns D and G if you inserted two columns in problem 7.2). The "t0 green bkg" and "t0 red bkg" columns give measurements of fluorescence from the local area around each spot on the filter; thus they give an indication of nonspecific binding of labeled RNA. We would like to exclude this from the intensity of each spot, as nonspecific binding may differ across the filter or may differ between the green and red labels.

- Insert a column following column D ("t0 green bkg"). Use the = sign (Edit Formula) button to subtract "t0 green bkg" from "t0 green." Make sure that you subtract the raw "t0 green" values, not the log values. Repeat for "t0 red bkg" and "t0 red."
- Create two additional columns and calculate log intensities for the background-adjusted intensities that you just created.
- Make a graph using these new adjusted, log 10 values. If background intensities were uniform across the array and between the red and green values, this graph should not be much different than the previous graph (see Fig. 7.3*c*).

[7-6] *Normalizing to total fluorescence.* Because there may have been more RNA in one sample than in the other, it is important to normalize each individual intensity value to the average fluorescence in that sample.

- Insert one new column in which you will calculate normalized raw minus background values for green data and one for red data. (*Note:* We will be normalizing the values in the raw minus background columns, not the log values that you used for the last graph. If you have done all of the steps so far, these should be columns E and J.)
- Find the average (mean) value for each of the raw minus background columns. To do this, click on the space below the last value in the column (e.g., box E6120). Click the = sign and type in AVERAGE(E2:E6119). When you hit Enter, the mean of the entire column will appear in the box. You will then divide each of the values in the raw minus background column by this mean. To do this, click on box 2 in your new column, click the = sign and type in E2/xxxx, where xxxx is the numerical value (not the box location on the spreadsheet) of the mean of the raw minus background column. Copy and paste this formula into the rest of the column. Repeat for the red data.
- Make two new columns, one for green and one for red. In these columns calculate the log of the normalized

raw minus background values (i.e., the columns that you made in the previous step with (raw minus normalized) divided by (mean of raw minus normalized)). Make a graph with these new columns.

[7-7] *Plotting log ratio of intensities vs. mean log intensity.* It is useful to look at the ratio of intensities as a function of mean log intensity because there may be more variation in expression levels of genes that are expressed at a low level than genes whose transcripts are abundant (or vice versa). This plot is also helpful in detecting curvature in the cloud of data points.

- Create a total of three new columns in the spreadsheet. In one column, calculate for each gene the average of the last two columns that you created in problem 7.6. These numbers will represent the green–red average of the log of the normalized background-subtracted intensities for each gene. This column will be the x axis of the graph.
- In another column, divide each value in the red normalized raw minus background column by each value in the green normalized raw minus background column. Do not use the log-normalized columns (i.e., the last columns that you created in problem 7.6), but use the columns that you created before taking the log in problem 7.6.
- In the third column, take the log of the values in the column that you just created. So, this column will be the log of the ratio of normalized background-subtracted intensities. Use this column as the y axis of this graph.
- The graph will be similar to the one in Figs. 7.4 and 7.5. Mouse over any outliers to learn their identities.

SELF-TEST QUIZ

[7-1] It is necessary to normalize microarray data because:
(a) Gene expression values are not normally distributed.
(b) Some experiments use cDNA labeled with fluorescence while others employ cDNA labeled with radioactivity.
(c) The efficiency of dye incorporation or radioactivity incorporation may vary for different samples.
(d) The efficiency of dye incorporation may vary for different samples.

[7-2] Microarray data analysis can be performed with scatter plots. The information you get from a scatter plot includes all of the following EXCEPT:
(a) You can tell whether a gene is expressed at a relatively high level or a low level.
(b) You can tell whether a gene has been up regulated or down regulated.
(c) You can tell whether a gene forms a cluster with other genes on the microarray.
(d) You can tell whether a gene is among the 5% most regulated genes in that experiment.

[7-3] Log ratios of gene expression values are often used rather than raw ratios because:
(a) Twofold up regulation or twofold down regulation log ratios each have the same absolute value.
(b) Twofold up regulation or twofold down regulation log ratios each have the same relative value.
(c) The scale of log ratios is compressed relative to the scale of raw ratios.
(d) A plot of log ratios compresses the expression values.

[7-4] Inferential statistics can be applied to microarray data sets to perform hypothesis testing:
(a) In which the probability is assessed that any individual gene is significantly regulated in a comparison of two samples
(b) In which the probability is assessed that any individual gene is significantly regulated in a comparison of two or more samples
(c) By clustering of array data
(d) By either supervised or unsupervised analyses

[7-5] Which one of the following statements is FALSE?
(a) Clustering of microarray data produces a tree that can resemble a phylogenetic tree.
(b) Clustering of microarray data can be performed on genes and/or samples.
(c) Clustering of microarray data can be performed with partitioning methods (such as k-means) or hierarchical methods (such as agglomerative or divisive clustering).
(d) Clustering of microarray data is always performed using principal components analysis.

[7-6] Clustering techniques rely on distance metrics to:
(a) Describe whether a clustering tree is agglomerative or divisive
(b) Reduce the dimensionality of a highly dimensional data set
(c) Identify the absolute values of gene expression measurements in a matrix of gene expression values versus samples
(d) Define the relatedness of gene expression values from a matrix of gene expression values versus samples

[7-7] A self-organizing map:
(a) Imposes some structure on the formation of clusters
(b) Is unstructured, like k-means clustering
(c) Has neighboring nodes that represent dissimilar clusters
(d) Cannot be represented as a clustering tree

[7-8] Principal components analysis (PCA):
(a) Minimizes entropy to visualize the relationships among genes and proteins
(b) Can be applied to gene expression data from microarrays but not to protein analyses

(c) Can be performed by agglomerative or divisive strategies
(d) Reduces highly dimensional data to show the relationships among genes or among samples

[7-9] The main difference between supervised and unsupervised analyses of microarray data is:
(a) Supervised approaches assign some prior knowledge about function to the genes and/or samples, while unsupervised analyses do not.
(b) Supervised approaches assign a fixed number of clusters, while unsupervised analyses do not.
(c) Supervised approaches cluster genes and/or samples, while unsupervised approaches cluster only genes.
(d) Supervised approaches include algorithms such as support vector machines and decision trees, while unsupervised approaches use clustering algorithms.

SUGGESTED READING

For cluster analysis of microarray data, Michael Eisen and colleagues (1998) describe the clustering of 8600 human genes as a function of time. This classic paper includes an excellent description of the metric used to define the relationships of gene expression values and also a discussion of the usefulness of clustering in defining functional relationships among expressed genes.

REFERENCES

Alizadeh, A. A., et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403,** 503–511 (2000).

Alizadeh, A. A., Ross, D. T., Perou, C. M., and van de Rijn, M. Towards a novel classification of human malignancies based on gene expression patterns. *J. Pathol.* **195,** 41–52 (2001).

Alter, O., Brown, P. O., and Botstein, D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA* **97,** 10101–10106 (2000).

Ball, C. A., et al. An open letter to the scientific journals. *Bioinformatics* **18,** 1409 (2002a).

Ball, C. A., et al. Standards for microarray data. *Science* **298,** 539 (2002b).

Beissbarth, T., et al. Processing and quality control of DNA array hybridization data. *Bioinformatics* **16,** 1014–1022 (2000).

Bouton, C. M., Hossain, M. A., Frelin, L. P., Laterra, J., and Pevsner, J. Microarray analysis of differential gene expression in lead-exposed astrocytes. *Toxicol. Appl. Pharmacol.* **176,** 34–53 (2001).

Bouton, C. M., and Pevsner, J. DRAGON View: Information Visualization for Annotated Microarray Data. *Bioinformatics* **18,** 323–324 (2002).

Bouton, C. M., and Pevsner, J. DRAGON: Database Referencing of Array Genes Online. *Bioinformatics* **16,** 1038–1039 (2000).

Bouton, C. M., Henry, G., Colantuoni, C., and Pevsner, J. DRAGON and DRAGON view: methods for the annotation, analysis, and visualization of large-scale gene expression data. In: G. Parmigiani, E. S. Garrett, R. A. Irizarry, and S. L. Zeger (eds.), *The Analysis of Gene Expression Data: Methods and Software.* Springer, New York, 2003, pp. 185–209.

Brazma, A. On the importance of standardisation in life sciences. *Bioinformatics* **17,** 113–114 (2001).

Brazma, A., Robinson, A., Cameron, G., and Ashburner, M. One-stop shop for microarray data. *Nature* **403,** 699–700 (2000).

Brazma, A., et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* **29,** 365–371 (2001).

Brazma, A., and Vilo, J. Gene expression data analysis. *FEBS Lett.* **480,** 17–24 (2000).

Brown, M. P., et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA* **97,** 262–267 (2000).

Chu, S., et al. The transcriptional program of sporulation in budding yeast. *Science* **282,** 699–705 (1998).

Claverie, J. M. Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* **8,** 1821–1832 (1999).

Colantuoni, C., Henry, G., Zeger, S., and Pevsner, J. SNOMAD (Standardization and NOrmalization of MicroArray Data): Web-accessible gene expression data analysis. *Bioinformatics* **18,** 1540–1541 (2002a).

Colantuoni, C., Henry, G., Zeger, S., and Pevsner, J. Local mean normalization of microarray element signal intensities across an array surface: Quality control and correction of spatially systematic artifacts. *Biotechniques* **32,** 1316–1320 (2002b).

Dopazo, J., Zanders, E., Dragoni, I., Amphlett, G., and Falciani, F. Methods and approaches in the analysis of gene expression data. *J. Immunol. Methods* **250,** 93–112 (2001).

Eickhoff, B., Korn, B., Schick, M., Poustka, A., and van der Bosch, J. Normalization of array hybridization experiments in differential gene expression analysis. *Nucleic Acids Res.* **27,** e33 (1999).

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95,** 14863–14868 (1998).

Golub, T. R., et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537 (1999).

Haverty, P. M., et al. HugeIndex: A database with visualization tools for high-density oligonucleotide array data from normal human tissues. *Nucleic Acids Res.* **30,** 214–217 (2002).

Hegde, P., et al. A concise guide to cDNA microarray analysis. *Biotechniques* **29,** 548–550, 552–554, 556 passim (2000).

Hilsenbeck, S. G., et al. Statistical analysis of array expression data as applied to the problem of tamoxifen resistance. *J. Natl. Cancer. Inst.* **91,** 453–459 (1999).

Iyer, V. R., et al. The transcriptional program in the response of human fibroblasts to serum. *Science* **283,** 83–87 (1999).

Kaufman, L., and Rousseeuw, P. J. *FINDING GROUPS IN DATA. An Introduction to Cluster Analysis,* Wiley, New York, 1990.

Khan J., et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58,** 5009–5013 (1998).

Kuruvilla, F. G., Park, P. J., and Schreiber, S. L. Vector algebra in the analysis of genome-wide expression data. *Genome Biol.* **3,** 11.1–11.11 (2002).

Landgrebe, J., Wurst, W., and Welzl, G. Permutation-validated principal components analysis of microarray data. *Genome Biol.* **3,** 19.1–19.11 (2002).

Liao, B., Hale, W., Epstein, C. B., Butow, R. A., and Garner, H. R. MAD: A suite of tools for microarray data management and processing. *Bioinformatics* **16,** 946–947 (2000).

Manduchi, E., et al. Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics* **16,** 685–698 (2000).

Misra, J., et al. Interactive exploration of microarray gene expression patterns in a reduced dimensional space. *Genome Res.* **12,** 1112–1120 (2002).

Motulsky, A. G. Jewish diseases and origins. *Nat. Genet.* **9,** 99–101 (1995).

Olshen, A. B., and Jain, A. N. Deriving quantitative conclusions from microarray expression data. *Bioinformatics* **18,** 961–970 (2002).

Perou, C. M., et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA* **96,** 9212–9217 (1999).

Perou, C. M., et al. Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).

Quackenbush, J. Computational analysis of microarray data. *Nat. Rev. Genet.* **2,** 418–427 (2001).

Raychaudhuri, S., Stuart, J. M., and Altman, R. B. Principal components analysis to summarize microarray experiments: Application to sporulation time series. *Pac. Symp. Biocomput.* 455–466 (2000).

Raychaudhuri, S., Sutphin, P. D., Chang, J. T., and Altman, R. B. Basic microarray analysis: Grouping and feature reduction. *Trends Biotechnol.* **19,** 189–193 (2001).

Sawitzki, G. Quality control and early diagnostics for cDNA microarrays. *R. News* **2,** 6–10 (2002).

Schuchhardt, J., et al. Normalization strategies for cDNA microarrays. *Nucleic Acids Res.* **28,** E47 (2000).

Sherlock, G. Analysis of large-scale gene expression data. *Brief. Bioinform.* **2,** 350–362 (2001).

Shipp, M. A., et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **8,** 68–74 (2002).

Singer, C. The Fasciculo di Medicina Venice 1493. R. Lier and Co., Florence, 1925.

Smid-Koopman, E., et al. Gene expression profiles of human endometrial cancer samples using a cDNA-expression array technique: Assessment of an analysis method. *Br. J. Cancer* **83,** 246–251 (2000).

Spellman, P. T., et al. Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol.* **3,** 46.1–46.9 (2002).

Tamayo, P., et al. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96,** 2907–2912 (1999).

Tusher, V. G., Tibshirani, R., and Chu, G. Significance analysis of microarray applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA* **98,** 5116–5121 (2001).

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. Systematic determination of genetic network architecture. *Nat. Genet.* **22,** 281–285 (1999).

Wall, M. E., Dyck, P. A., and Brettin, T. S. SVDMAN—singular value decomposition analysis of microarray data. *Bioinformatics* **17,** 566–568 (2001).

West, M., et al. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl. Acad. Sci. USA* **98,** 11462–11467 (2001).

Wittes, J., and Friedman, H. P. Searching for evidence of altered gene expression: A comment on statistical analysis of microarray data. *J. Natl. Cancer. Inst.* **91,** 400–401 (1999).

Zolman, J. F. *Biostatistics.* Oxford University Press, New York, 1993.

	1	2	3	4	5	6	7
	Globin of the Oxyhæmoglobin of Horse's Blood.	Serum-albumin of Horse's Blood.	Serum-globulin.	Egg-white.	Egg-albumin crystallised.	Albumin of Yolk.	Caseinogen (see p. 397 for other products).
Glycocoll	0 [9]	0 [10]	3·52 [95]	...	...	...	0 [45 95]
Alanin	4·19 [9]	2·68 [10]	2·22 [95]	...	...	...	0·9 [95]
Leucin	29·04 [9]	20·48 [10]	18·7 [95]	22·6 [40]	...	...	10·5 [95]
Phenylalanin	4·24 [9]	3·08 [10]	3·84 [95]	+ [3]	...	...	3·2 [95]
α-Pyrrolidin-carboxylic acid	2·34 [9]	1·04 [10]	2·76 [95]	+ [3]	...	...	3·2 [1 95]
Glutaminic acid	1·73 [9]	1·52 [10]	2·20 [95]	+ [40]	+ [48]	...	10·7 [95]
Aspartic acid	4·43 [9]	3·12 [10]	2·54 [95]	+ [37]	+ [48]	...	1·2 [95]
Cystin	0·31 [9]	2·53 [43]	1·51 [43]	0·4 [43]	0·29 [43]	...	0·065
Serin	0·56 [9]	0·6 [10]	...	...	...	...	0·43 [14]
Oxy-α-Pyrrolidin-carboxylic acid	1·04 [9]	...	...	...	...	...	0·25 [14]
Tyrosin	1·33 [9]	2·1 [10]	...	0·58 [66]	1·5 [31]	...	4·5 [36 66]
Lysin	4·28 [9]	...	...	+ [16 75]	...	...	5·8 [26]
Histidin	10·96 [9]	..	...	+ [18]	+ [18]	+ [18]	2·6 [26]
Arginin	5·42 [9]	...	...	+ [17]	+ [16 17]	+ [17]	4·84 [26]
Tryptophane	+ [9]	+ [10]	...	+ [80]	...	...	1·5 [54]
Ammonia	0·93 [50]	1·2 [49]	1·75 [49]	...	1·5 [49]	...	1·8 [26]
Cystein	...	+ [44]	...	+ [44]	...	...	0 [44]
Amino-valerianic acid	...	...	...	...	...	...	1·95
Glucosamin	...	...	...	...	10–11 [34]	...	0 [67]
Diamino-trioxydodecanoic acid	...	...	...	...	...	...	0·75 [92]

While it is obvious to us that most proteins are composed of twenty amino acids, chemists in the late 19th century struggled to understand protein composition. At the turn of the century only several dozen proteins were known, including so-called albumins (including serum albumins, lactoglobulins, fibrinogen, myosin, and histones), proteids (e.g. hemoglobin and mucins), and albuminoids (e.g. collagen, keratin, elastin, and amyloid). Of these proteins only a very small group were available in pure form as crystals (e.g. hemoglobin and serum albumin from horse, ovalbumin, and ichthulin [salmon albumin]). Gustav Mann (1906, p. 70–75) described the dissociation products of 51 assorted proteins into their fundamental units. The results are shown for seven proteins (see columns). The rows indicate various compounds found upon dissolving the proteins. Most of these are amino acids; for example, glycocoll is a name formerly given to glycine. This table shows that from the earliest times that proteins could be analyzed, scientists made an effort to understand both the nature of individual protein molecules and the relationships of related proteins from different species.

8

Protein Analysis and Proteomics

INTRODUCTION

A living organism consists primarily of five substances: proteins, nucleic acids, lipids, water, and carbohydrates. Of these essential ingredients, it is the proteins that most define the character of each cell. DNA has often been described as a substance that corresponds to the blueprints of a house, specifying the materials used to build the house. These materials are the proteins, and they perform an astonishing range of biological functions. This includes structural roles (e.g., actin contributes to the cytoskeleton), roles as enzymes (proteins that catalyze biochemical reactions, typically increasing a reaction rate by several orders of magnitude), and roles in transport of materials within and between cells. If DNA is the blueprint of the house, proteins form primary components not just of the walls and floors of the house but also of the plumbing system, the system for generating and transmitting electricity, and the trash removal system.

Proteins are polypeptide polymers consisting of a linear arrangement of amino acids. The word *protein* was coined in the 1830s by the Swedish chemist Jöns Jacob Berzelius (1779–1848). The Greek word *proteios* means "of first rank."

APPROACHES TO PROTEINS: FOUR PERSPECTIVES

This chapter begins with four different perspectives on proteins (summarized in Fig. 8.1):

1. Protein families (domains and motifs)
2. Physical properties of proteins

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

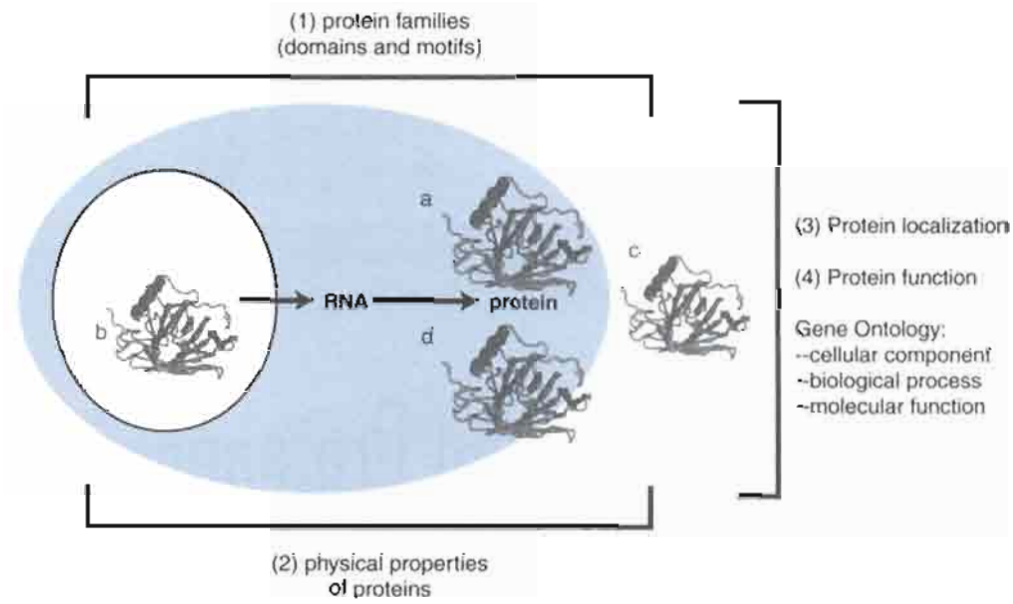


FIGURE 8.1. Overview of proteins. A protein is composed of a series of amino acids specified by a gene. Proteins can be classified by a variety of criteria, including family, localization, physical properties, and function. (1) Protein families are defined by the homology of a protein to other proteins; the proteins may be homologous over a partial region. Databases of protein families and motifs (discussed in Chapter 10) allow hundreds of thousands of proteins to be classified in groups that may be functionally related. (2) Proteins may be described in terms of their physical properties, such as size (molecular weight), shape (e.g., Stokes radius and frictional coefficient), charge (isoelectric point), posttranslational modifications (see text), or the existence of isoforms due to proteolytic processing or alternative mRNA splicing. (3) The Gene Ontology (GO) Consortium classifies proteins according to cellular component (i.e., localization), biological process (e.g., transcription or endocytosis), and molecular function (e.g., enzyme or transporter). A protein can belong to multiple categories of any of these groups. The GO system provides a dynamic, controlled vocabulary that can be applied to all eukaryotic proteins. In this figure, a protein is depicted in several possible locations: it may be soluble in the cytosol (label a), in an intracellular organelle such as the nucleus (b), or extracellular as a secreted protein (c). A protein may be bound to membranes on the cell surface (d) or on an intracellular organelle (not shown); membrane localization may be via transmembrane domains or by peripheral attachment.

3. Protein localization
4. Protein function

The first perspective we will consider is the protein family. We will define terms such as family, domain, and motif. Next, we will consider the physical properties of proteins and how we can assess them.

The next ways to consider proteins correspond in part to a conceptual framework provided by the Gene Ontology (GO) Consortium. In the past few years billions of base pairs of DNA have been sequenced, including the complete genomes of dozens of organisms (Chapters 12–17). A major challenge to the field of bioinformatics is to identify protein-coding genes (see Chapter 12). Another great challenge is to annotate them, that is, to provide a description of their nature and function. The GO Consortium (Ashburner et al., 2000) provides a flexible, controlled vocabulary to describe three aspects of proteins: cellular component, biological process, and molecular function. We will provide an overview of the GO project. We will then provide a general description of the protein localization and protein function, corresponding to the GO categories of cellular component and molecular function.

In this chapter, after discussing four perspectives on individual proteins, we will explore approaches to studying the complete collection of proteins encoded by a genome (the "proteome") and the ways in which proteins interact with each other in

pathways. In Chapter 9, we will consider the structure of proteins. In Chapter 10 we will learn how to multiply align homologous proteins (and nucleic acids). We will also address databases of aligned proteins, such as the Protein Families Database (Pfam) and InterPro. We will visualize multiple sequence alignments as phylogenetic trees in Chapter 11.

Perspective 1. Protein Domains and Motifs: Modular Nature of Proteins

Let us begin our discussion of protein domains by considering several types of proteins. In the simplest case, a protein (or gene) has no matches to any other sequences in the available databases. This situation occurs quite frequently, and it is not unusual to find that half the predicted proteins in a completed genome have no identifiable homologs (see Chapters 12–17). Even if there are no known homologs, a protein may have features such as a transmembrane domain, potential sites for phosphorylation, or some predicted secondary structure (see below and Chapter 9). Such features may give clues to the structure and/or function of the protein.

For proteins that do have orthologs and/or paralogs, there are regions of significant amino acid identity between at least two proteins (or DNA sequences). Such regions of proteins that share significant structural features and/or sequence identity have a variety of names: signatures, domains, modules, modular elements, folds, motifs, patterns, or repeats. These terms have varied definitions, but all refer to the idea that there are closely related amino acid sequences shared by multiple proteins (Bork and Gibson, 1996). We will primarily adopt the definitions provided by the InterPro Consortium (Mulder et al., 2002, 2003). InterPro is an integrated documentation resource that encompasses a group of databases of protein families, domains, and functional sites.

InterPro and its member databases (PRINTS, PROSITE, Pfam, ProDom, SMART, and TIGRFAMs) will be described in this chapter and Chapter 10. InterPro is accessed at ▸http://www.ebi.ac.uk/interpro/.

A *signature* is a broad term that denotes a protein category, such as a domain or family or motif (see below). When you consider a single protein sequence in isolation, there is only a limited amount of information you can infer about its structure or function. However, when you align related sequences, a consensus sequence may be identified. There are two principal kinds of signatures, and each is identified with its own methodology.

1. A domain is a region of a protein that can adopt a particular three-dimensional structure (Doolittle, 1995). Domains are also called modules (Henikoff et al., 1997; Sonnhammer and Kahn, 1994). The term *fold* is commonly used in the context of three-dimensional structure (Jones, 2001). Together, a group of proteins that share a domain is called a family. Many protein domains are further classified based upon the subcellular localization of the domain (e.g., intracellular domains of proteins occur in the cytoplasm; extracellular domains are oriented outside the cell) or in terms of the structure of the domain (e.g., zinc finger domains bind the divalent cation zinc).

There are many databases of protein families, such as Pfam and SMART, that we will explore in Chapter 10. The definitions of the terms *family, domain, repeat,* and related terms in the InterPro and SMART databases are given in Tables 8.1 and 8.2.

2. Motifs (or fingerprints) are short, conserved regions of proteins (discussed below). A motif typically consists of a pattern of amino acids that characterizes a protein family (Bork and Gibson, 1996). The size of a defined motif is often 10–20 contiguous amino acid residues, although it can be smaller or larger. Some simple and common motifs, such as a stretch of amino acids that form a transmembrane domain or a consensus phosphorylation site, do not imply homology when found

TABLE 8-1 Definitions from InterPro Database of Protein Families and Related Terms

Term	Definition
Family	An InterPro family is a group of evolutionarily related proteins that share one or more domains/repeats in common. A InterPro entry of "type = family" may contain a signature for a small conserved region that is representative of the family and therefore need not necessarily cover the whole protein.
Domain	An InterPro domain is an independent structural unit which can be found alone or in conjunction with other domains or repeats. Domains are evolutionarily related. An InterPro entry of "type = domain" is diagnostic for a domain but does not necessarily define domain boundaries exactly.
Repeat	An InterPro repeat is a region that is not expected to fold into a globular domain on its own. For example, six to eight copies of the *WD40* repeat are needed to form a single globular domain. There also many other short repeat motifs that probably do not form a globular fold that have "type = repeat."
Posttranslational modification	A posttranslational modification includes, for example, an N-glycosylation site. The sequence motif is defined by the molecular recognition of this region in a cell. This may group together proteins that need not be evolutionarily related.

Source: Adapted from http://www.ebi.ac.uk/interpro/user_manual.html.

in a group of proteins. In other cases a small motif may provide a characteristic signature for a protein family.

To introduce specific examples of domains, Table 8.3 lists the most common domains in the proteins encoded by the human genome. Similar lists are available for the abundant protein domains of other organisms (Chapters 14–16). In many cases, two proteins that share a domain also share a common function. For example, of the hundreds of small guanosine triphosphate- (GTP-) binding proteins

TABLE 8-2 Definitions of Protein Domains and Motifs from SMART Database

SMART is a tool to allow automatic identification and annotation of domains in user-supplied protein sequences (see Chapter 10).

Term	Definition
Domain	Conserved structural entities with distinctive secondary structure content and a hydrophobic core. In small disulfide-rich and Zn^{2+}-binding or Ca^{2+}-binding domains, the hydrophobic core may be provided by cystines and metal ions, respectively. Homologous domains with common functions usually show sequence similarities.
Domain composition	Proteins with the same domain composition have at least one copy of each domain of the query.
Domain organization	Proteins having all the domains as the query in the same order (additional domains are allowed).
Motif	Sequence motifs are short conserved regions of polypeptides. Sets of sequence motifs need not necessarily represent homologs.
Profile	A profile is a table of position-specific scores and gap penalties, representing an homologous family that may be used to search sequence databases (Bork and Gibson, 1996).

Source: Adapted from http://smart.embl-heidelberg.de/help/smart_glossary.shtml.

TABLE 8-3 Fifteen Most Common Domains of *Homo sapiens*

InterPro ID	Matches per Genome	Number of Proteins	Name
IPR000822	30034	1093	Zn finger, C2H2 type
IPR003006	2631	1032	Immunoglobulin/major histocompatibility complex
IPR000561	4985	471	EGF-like domain
IPR001841	1356	458	Zn-finger, RING
IPR001356	2542	417	Homeobox
IPR001849	1236	405	Pleckstrin-like
IPR000504	2046	400	RNA-binding region RNP-1 (RNA recognition motif)
IPR001452	2562	394	SH3 domain
IPR002048	2518	392	Calcium-binding EF-hand
IPR003961	2199	300	Fibronectin, type III
IPR001478	1398	280	PDZ/DHR/GLGF domain
IPR005225	261	261	Small GTP-binding protein domain
IPR000210	583	236	BTB/POZ domain
IPR001092	713	226	Basic helix-loop-helix (bHLH) dimerization domain
IPR002126	5168	226	Cadherin

Source: From the European Bioinformatics Institute (EBI) proteome analysis site (▸http://www.ebi.ac.uk/proteome/) (August 2002) based upon the InterPro database (▸http://www.ebi.ac.uk/interpro/index.html).

(Table 8.3), many dozens are thought to regulate the intracellular docking and fusion of transport vesicles through a cycle of GTP binding and hydrolysis (Geppert et al., 1997). Other, related low-molecular-weight GTP-binding proteins function in cell cycle control and cytoskeletal organization (reviewed in Takai et al., 2001). This superfamily is organized into related subfamilies that are usually presumed to share common functions.

Focusing our attention on a single domain, there are many ways in which proteins can share that domain in common (Fig. 8.2). The entire protein may consist of one domain, such as the lipocalin domain (Fig. 8.2*a*). Many other small, globular proteins also consist of a single domain.

It is even more common for a domain to form a subset of a protein. A comparison of two proteins often indicates that the domains occupy different regions of each protein (Fig. 8.2*b*). A group of six proteins contain a domain that confers the ability of each protein to bind methylated DNA. One of these proteins, methyl-CpG-binding protein 2 (MeCP2), is a transcriptional repressor that binds the regulatory region of a variety of genes. (Mutations in the *MECP2* gene cause Rett syndrome, a neurological disorder that affects girls and is one of the most common causes of mental retardation in females. See box 18.2.) We can perform a blastp search with the MeCP2 protein sequence to illustrate the concept of protein domains. The BLAST formatting page shows that the methyl-CpG-binding domain (MBD) is present in several databases of protein domains (Fig. 8.3*a*). The BLAST search result shows that a portion of MeCP2 matches five other MBD proteins (Fig. 8.3*b*). Furthermore, examination of the MeCP2/MBD family shows that the proteins are various different sizes, sharing in common only the MBD domain (Fig. 8.3*c*).

What is the definition of a family; is a group of proteins homologous if they share only one domain in common? The MBD domains are clearly homologous

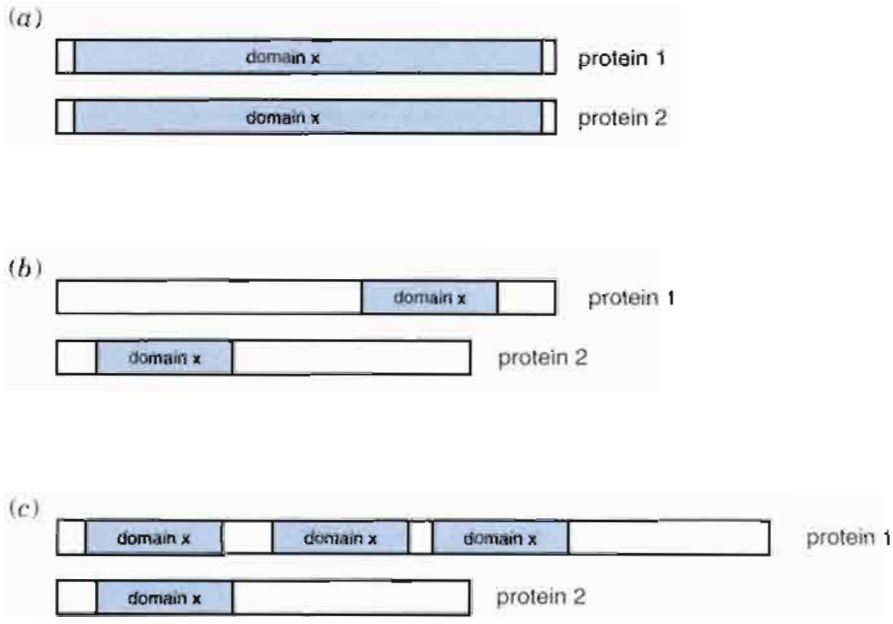


FIGURE 8.2. Proteins can share a common domain in a number of ways. (a) *A domain may extend essentially across the length of a protein. An example of this format is the lipocalin family.* (b) *Domains may contain highly related stretches of amino acids that form only a subset of each protein's sequence. An example of this situation is found in the family of transcriptional regulators that bind methylated DNA (see Fig. 8.3).* (c) *A domain may be repeated within a single protein (sometimes with many copies). Such a domain may occur in homologous proteins any number of times. An example is the family of proteins containing a fibronectin III–like repeat (Fig. 8.4).*

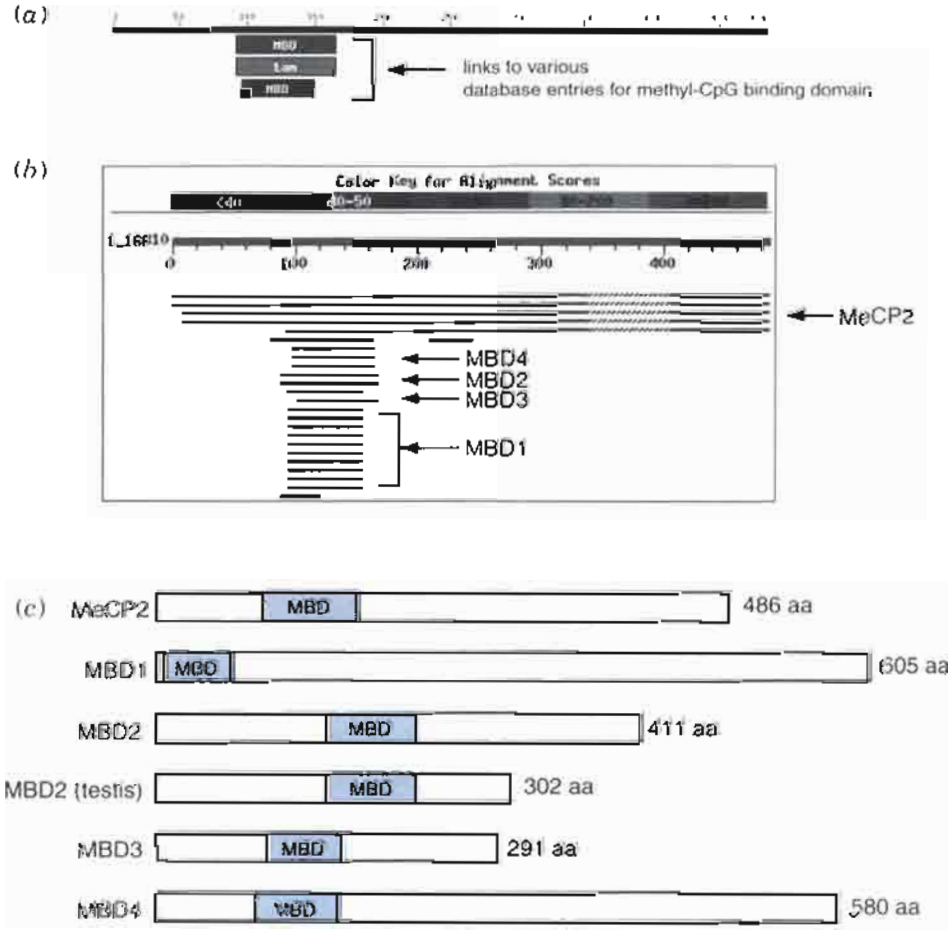


FIGURE 8.3. A methyl-binding domain is found in several human proteins. To illustrate the concept of domains, methyl-CpG-binding protein 2 (MeCP2; NP.004983) was used as a query in a blastp search restricted to human proteins. (a) *The formatting BLAST web page shows that this protein has a domain that is present in several databases.* (b) *The BLAST search reveals there are five separate MeCP2 entries that match the query (top five alignments). Additionally, there is a region of about 80 amino acids in MeCP2 that matches other methyl-CpG-binding proteins: MBD1 (NP.056671), MBD2 (NP.003918), a testis-specific isoform of MBD2 (NP.056647), MBD3 (NP.003917), and MBD4 (NP.003916).* (c) *These proteins have different sizes. Also, the methylated DNA-binding domain that these proteins share occurs in different regions of the proteins. Further BLAST searches confirm that together these six proteins share no significant amino acid identity at any region other than the methyl-binding domain.*

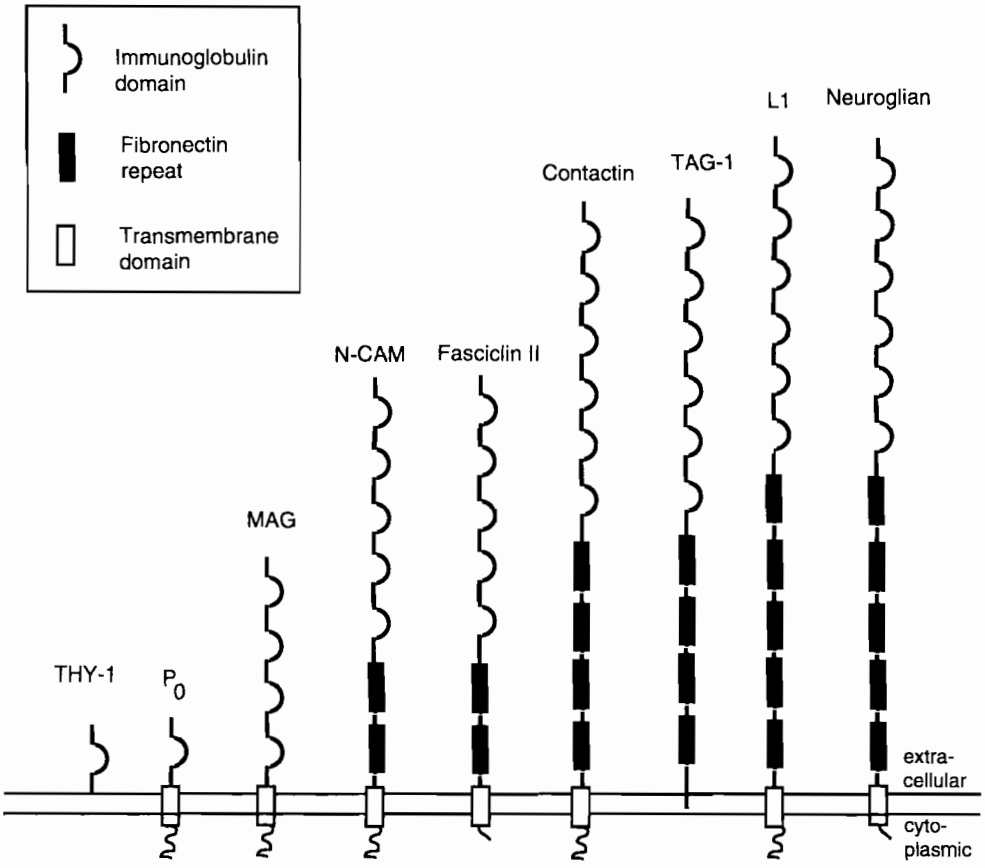


FIGURE 8.4. *Many proteins have multiple copies of distinct domains. Two of the most common domains are the immunoglobulin (Ig) domain and the fibronectin repeat. These domains are especially common in the extracellular domains of proteins.*

(descended from a common ancestor), defining this group of proteins as a family. But the regions outside the MBD domain share no significant amino acid identity. A family is a group of evolutionarily related proteins that share one (or more) regions of homology.

A third scenario for proteins containing individual domains is that the domain may be repeated many times (Fig. 8.2*c*). Two of the most common protein domains in *H. sapiens* are immunoglobulin domains and fibronectin repeats (Table 8.3). Both of these domains are present in variable numbers in a group of proteins having extracellular domains (Fig. 8.4). Notably, these and other extracellular domains are highly abundant in humans and the multicellular nematode *Caenorhabditis elegans* but nearly absent in the single-celled eukaryote *Saccharomyces cerevisiae* (Copley et al., 1999). Comparison of protein families that are encoded by various genomes sheds light on the biological processes that each organism performs (Chapters 12–17).

Added Complexity of Multidomain Proteins

So far we have focused on the subject of single domains. Multidomain proteins provide a common, more complicated scenario. HIV-1 pol is an example of such a protein (Frankel and Young, 1998) (see Fig. 8.7 below). The *pol* gene encodes a single large polypeptide that is cleaved into three independent proteins with distinct biochemical activities: an aspartyl protease, a reverse transcriptase (RNA-dependent

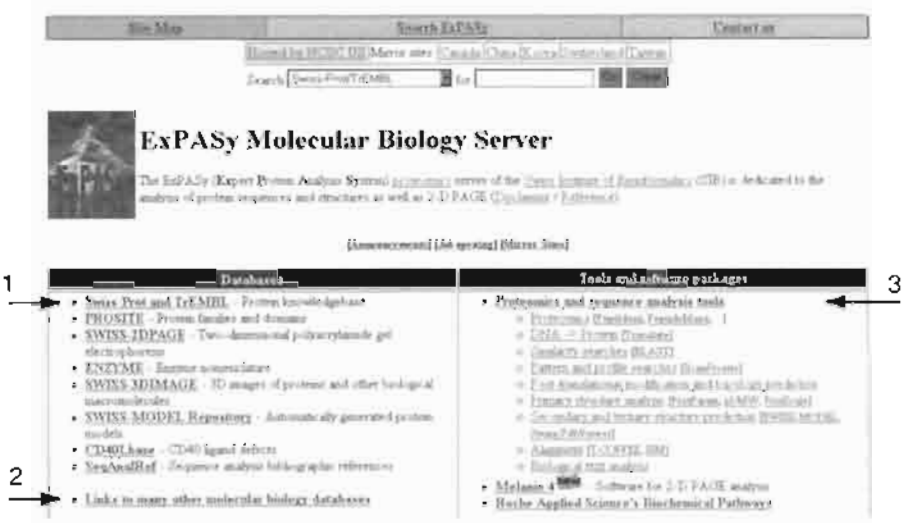


FIGURE 8.5. ExPASy offers the premier web server for protein analysis (http://www.expasy.ch/). The site provides a gateway to the sequence retrieveal system (arrow 1) and to a large, well-organized list of links to databases (arrow 2). A variety of tools for protein analysis are provided (arrow 3).

DNA polymerase), and an integrase. Note that HIV-1 pol presents an example of a multidomain protein that is cleaved into separate single-domain proteins. Other multidomain proteins, such as the immunoglobulin domain proteins depicted in Figure 8.4, maintain separate domains within a mature polypeptide.

To examine the sequence of pol, we will use the Expert Protein Analysis System (ExPASy) (Fig. 8.5). Go to the Sequence Retrieval System (SRS) and enter a search of HIV-1 (organism) and pol. Select the protein of 1003 amino acids (SwissProt accession P03369). This SwissProt record includes a variety of links to related databases (Fig. 8.6). Six of these links are to databases of protein families (or domains or

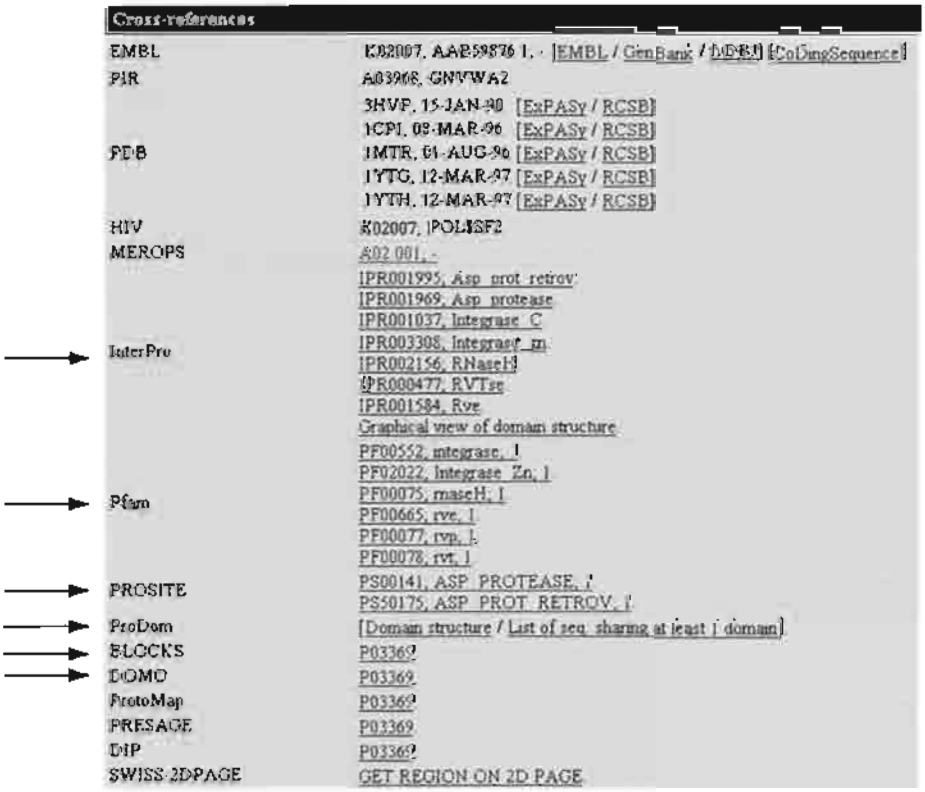


FIGURE 8.6. The SwissProt database entry for HIV-1 pol (accession P03369) includes many links to other databases of protein domains. In this chapter, we will explore some of these databases, including InterPro, Pfam, and PROSITE. This SwissProt entry was accessed by searching at the ExPASy Sequence Retrieval System (http://www.expasy.ch/srs5/). The arrows indicate six different databases with information on domains or motifs in HIV-1 pol.

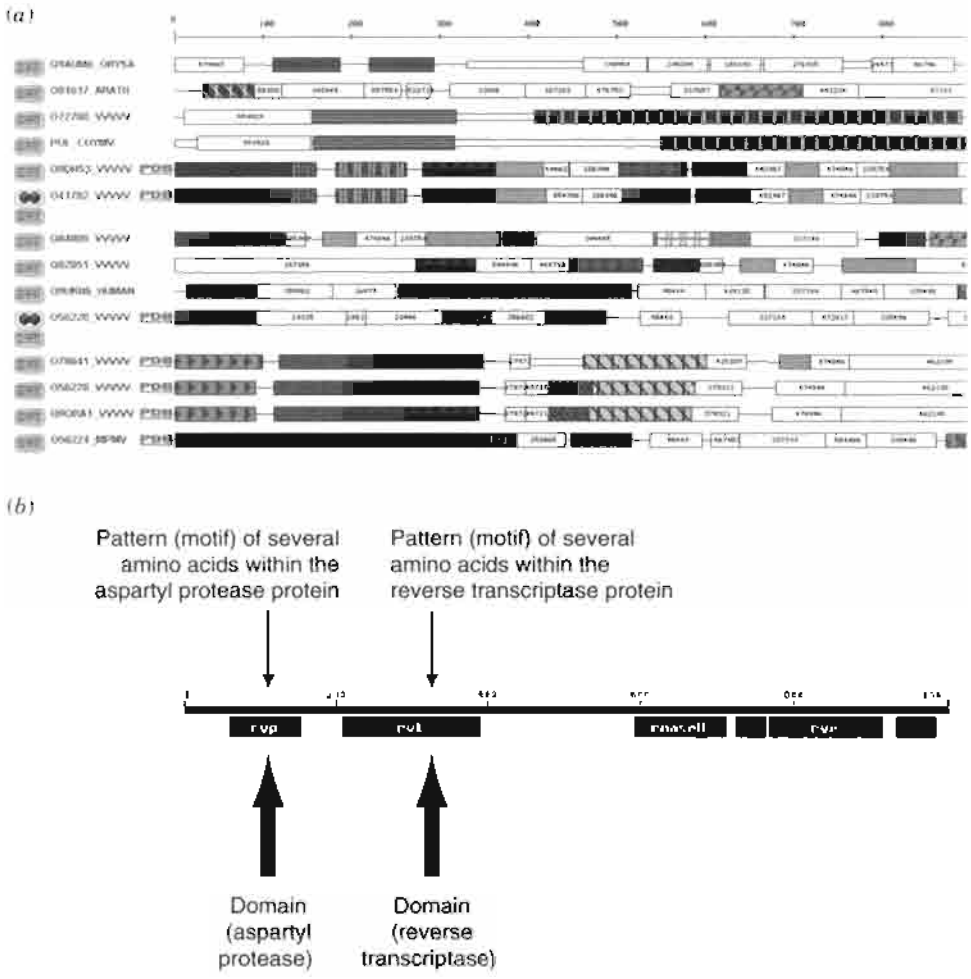


FIGURE 8.7. (a) *The ProDom link from SwissProt entry of HIV-1 pol shows hundreds of proteins that share domains in common with pol. This list is obtained by clicking the ProDom link ("List of seq. sharing at least 1 domain") from the SwissProt entry or it can be accessed directly from the ProDom database (▸http://prodes.toulouse.inra.fr/prodom/doc/prodom.html). Domains appear as distinct modules within protein sequences. A portion of fourteen proteins is shown here.* (b) *A protein may have domains (modules) which are relatively larger and patterns (motifs) which typically consist of only a few amino acids. Although a pattern or motif may not adopt a known three-dimensional structural conformation, it may nonetheless contain an amino acid sequence that is characteristic of a protein family.*

motifs): InterPro, Pfam, PROSITE, ProDom, BLOCKS, and DOMO (see below). Follow the ProDom link to a series of proteins sharing at least one domain in common with HIV-1 pol. The ProDom result is a graphical overview of hundreds of proteins that share regions in common with HIV-1 pol (Fig. 8.7).

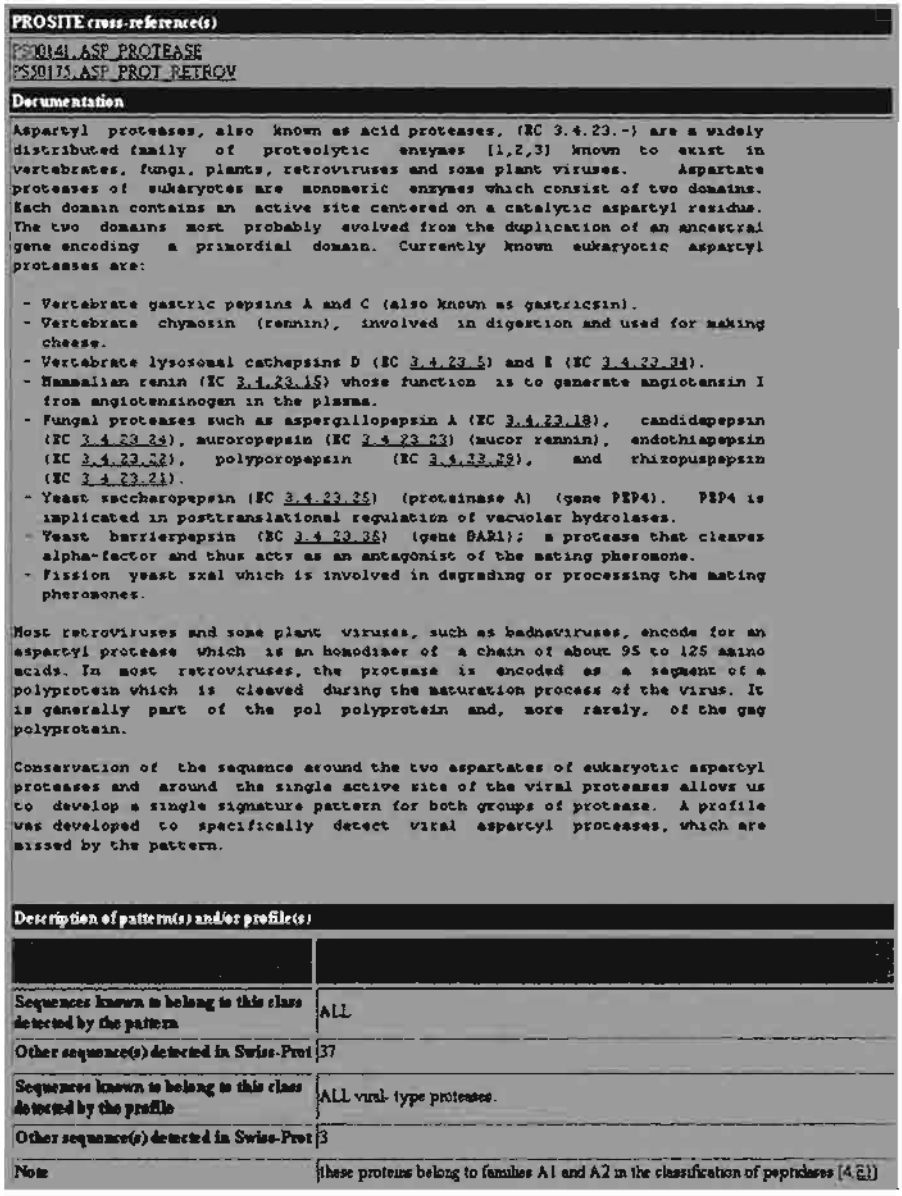
The ExPASy SRS is located at ▸http://www.expasy.ch/srs5/. Many of the tools we will explore in this chapter are available at ExPASy, which is part of the Swiss Institute of Bioinformatics.

Protein Patterns: Motifs or Fingerprints Characteristic of Proteins

Within a domain, there may be a small number of characteristic amino acid residues that occur consistently. These are called motifs (or fingerprints). An example of a motif is the amino acids that are reliably found at the active site of an enzyme. In the aspartyl protease domain of HIV-1 pol, an aspartate residue is crucial for the proteolytic reaction. PROSITE is a dictionary of protein motifs (Sigrist et al., 2002). Following the link from SwissProt (Fig. 8.6) or searching the site directly, one finds the entry for aspartyl proteases (Fig. 8.8). The motif is defined by a string of 12 amino acid residues: [LIVMFGAC]-[LIVMTADN]-[LIVFSA]-D-[ST]-G-[STAV]-[STAPDENQ]-x-[LIVMFSTNC]-x-[LIVMFGTA]. This format is identical to that used by PHI-BLAST (Chapter 5).

Motifs are subsets of protein domains. A short motif that is found in almost all lipocalins is GXW. The consensus pattern defined in PROSITE (document

PROSITE is accessed at ▸http://www.expasy.org/prosite/. In PROSITE, the term *profile* refers to a quantitative motif description based on a generalized profile syntax. The term *pattern* refers to a qualitative motif description based on a regular expression–like syntax such as those described below. The term *motif* refers to the biological object one attempts to approximate by a pattern or a profile. See ▸ftp://www.expasy.org/databases/prosite/release/profile.txt for these definitions.



PROSITE cross-reference(s)

PS00141, ASP_PROTEASE
PS50175, ASP_PROT_RETROV

Documentation

Aspartyl proteases, also known as acid proteases, (EC 3.4.23.-) are a widely distributed family of proteolytic enzymes [1,2,3] known to exist in vertebrates, fungi, plants, retroviruses and some plant viruses. Aspartate proteases of eukaryotes are monomeric enzymes which consist of two domains. Each domain contains an active site centered on a catalytic aspartyl residue. The two domains most probably evolved from the duplication of an ancestral gene encoding a primordial domain. Currently known eukaryotic aspartyl proteases are:

- Vertebrate gastric pepsins A and C (also known as gastricsin).
- Vertebrate chymosin (rennin), involved in digestion and used for making cheese.
- Vertebrate lysosomal cathepsins D (EC 3.4.23.5) and E (EC 3.4.23.34).
- Mammalian renin (EC 3.4.23.15) whose function is to generate angiotensin I from angiotensinogen in the plasma.
- Fungal proteases such as aspergillopepsin A (EC 3.4.23.18), candidapepsin (EC 3.4.23.24), auroropepsin (EC 3.4.23.23) (mucor rennin), endothiapepsin (EC 3.4.23.22), polyporopepsin (EC 3.4.23.29), and rhizopuspepsin (EC 3.4.23.21).
- Yeast saccharopepsin (EC 3.4.23.25) (proteinase A) (gene PEP4). PEP4 is implicated in posttranslational regulation of vacuolar hydrolases.
- Yeast barrierpepsin (EC 3.4.23.35) (gene BAR1); a protease that cleaves alpha-factor and thus acts as an antagonist of the mating pheromone.
- Fission yeast sxa1 which is involved in degrading or processing the mating pheromones.

Most retroviruses and some plant viruses, such as badnaviruses, encode for an aspartyl protease which is an homodimer of a chain of about 95 to 125 amino acids. In most retroviruses, the protease is encoded as a segment of a polyprotein which is cleaved during the maturation process of the virus. It is generally part of the pol polyprotein and, more rarely, of the gag polyprotein.

Conservation of the sequence around the two aspartates of eukaryotic aspartyl proteases and around the single active site of the viral proteases allows us to develop a single signature pattern for both groups of protease. A profile was developed to specifically detect viral aspartyl proteases, which are missed by the pattern.

Description of pattern(s) and/or profile(s)

Sequences known to belong to this class detected by the pattern	ALL
Other sequence(s) detected in Swiss-Prot	37
Sequences known to belong to this class detected by the profile	ALL viral-type proteases.
Other sequence(s) detected in Swiss-Prot	3
Note	these proteins belong to families A1 and A2 in the classification of peptidases [4,5])

FIGURE 8.8. PROSITE entry for aspartyl proteases, including HIV-1 pol. This database describes signatures (motifs) for over 1000 protein families.

PDOC00187) incorporates several additional amino acids surrounding GXW. That motif is [DENG]-x-[DENQGSTARK]-x(0,2)-[DENQARK]-[LIVFY]-{CP}-G-{C}-W-[FYWLRH]-x-[LIVMTA]. The GXW sequence is represented as G-{C}-W, where the curly brackets indicate that any amino acid other than cysteine is accepted at that position. Some motifs are extremely short and very common, such as the sequence surrounding a serine or threonine that is a substrate for many kinases. Such motifs are not specific to a particular protein family, and their occurrence in multiple proteins does not reflect homology. A search of PROSITE for "kinase" reveals three dozen entries, including both kinase and kinase substrate signatures. One of these entries is for the protein kinase C (PKC) consensus phosphorylation site, [ST]-x-[RK] (S or T is the phosphorylation site, and x is any residue) (PROSITE document PDOC00005). This simple motif occurs in proteins many thousands of times.

TABLE 8-4 Some Physical Properties of Proteins

Property	Classical Method	Example
Amino acid motifs	–	PDZ domain (e.g., nitric oxide synthase), coiled-coil domain (e.g., hemagglutinin, syntaxin, SNAP-25, myosin)
Isoelectric point (pI)	Derived from isoelectric focusing	–
Molecular weight	Derived from Stokes radius and sedimentation coefficient	–
Posttranslational modifications: phosphorylation	Enzymatic analyses	Synapsin
Posttranslational modifications: glycosylation	Enzymatic analyses	Nerve growth factor, neural cell adhesion molecule
Posttranslational modifications: isoprenylation	–	Lamin B, G protein γ subunits, *rab3A*
Posttranslational modifications: palmitoylation	–	β-Adrenergic receptor, *GAP-43,* insulin receptor, rhodopsin, nAChR
Posttranslational modifications: myristoylation	–	PKA, $G_{i\alpha}$-subunit, MARCKS protein, calcineurin
Posttranslational modifications: GPI-anchored proteins	Enzymatic analyses	Alkaline phosphatase, *thy-1,* prion protein, 5′-nucleotidase, uromodulin
Sedimentation coefficient	Derived from sucrose density gradients	–
Stokes radius	Derived from gel filtration	–
Transmembrane domain	Derived from subcellular fractionation	–

Abbreviations: G protein, guanosine triphosphate-binding protein; GAP-43, growth-associated protein of 43 kD; MARCKS, myristoylated alanine-rich C-kinase substrate; nAChR, nicotinic acetylcholine receptor; PDZ domain, *p*ost-synaptic density ptotein PSD-95, the *Drosophila* tumor suppressor *d*iscs-large, tight-junction protein ZO-1; PKA, protein kinase A; SNAP-25, synaptosomal-associated protein of 25 kD; Rab3A, rat brain GTP-binding protein 3A; thy-1, thymocyte-1.

An important aspect of regular expressions (or patterns) in the PROSITE database is that they are qualitative (i.e., either matching or not) and not quantitative. While patterns can accommodate complex definitions, such as having one of several different amino acid residues in a given position, mismatches are not tolerated when a protein sequence is compared to a pattern. In contrast to such rigid patterns, many databases such as Pfam, ProDom, SMART, and TIGRFAMs (described in Chapter 10) use profiles. Profiles, like patterns, are built from multiple sequence alignments, but they employ position-specific scoring matrices. They also span larger stretches of protein sequence than do patterns.

For websites offering protein motif analysis tools, see Table 8.11 under Web Resources.

Perspective 2. Physical Properties of Proteins

Proteins are characterized by a variety of physical properties that derive both from their essential nature as an amino acid polymer and from a variety of posttranslational modifications (Table 8.4). Some of these modifications allow the covalent attachment of a hydrophobic group to a protein to promote insertion into a lipid bilayer. Examples include palmitoylation, farnesylation, myristylation, and inositol glycolipid attachment (Fig. 8.9).

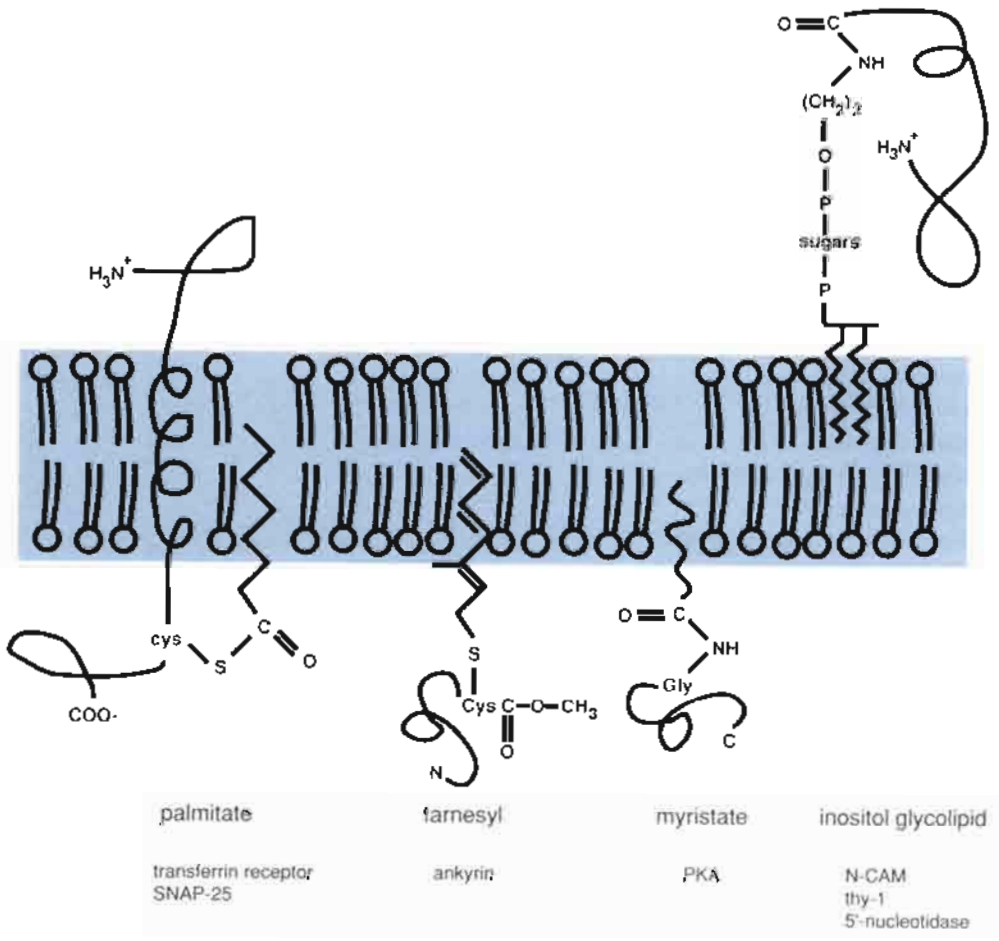


FIGURE 8.9. A variety of posttranslational modifications are added to proteins. Examples are palmitoylation (e.g., to the transferrin receptor and SNAP-25), farnesylation (e.g., to ankyrin), myristoylation (e.g., to protein kinase A), and inositol glycolipid anchoring to a membrane (e.g., neural cell adhesion molecule, thy-1, and 5'-nucleotidase). While these covalent modifications can be studied biochemically, a variety of websites offer predictions of possible sites of covalent modification to proteins. Adapted from Austen and Westwood (1991), p. 42. Used with permission.

A variety of web-based services are available to evaluate the predicted physical properties of proteins. Resources are available to input an individual protein sequence and to predict its physical properties, such as mass and isoelectric point (pI; Fig. 8.10 and Table 8.14 under Web Resources), amino acid composition, glycosylation sites (Table 8.15), phosphorylation sites (Table 8.16), and tyrosine sulfation (Fig. 8.11).

How can we assess the accuracy of these various prediction programs? In recent decades, the physical properties of proteins were assessed at the laboratory bench, one protein at a time (Cooper, 1977). The molecular mass of a protein can be estimated by gel filtration chromatography or by polyacrylamide gel electrophoresis (PAGE). Its shape can be estimated by calculating the frictional coefficient, obtained through a combination of gel filtration and sucrose density gradient centrifugation. Such techniques cannot be applied to large numbers of proteins. Almost all proteins that are studied using the tools of bioinformatics have not been purified, but instead the protein sequence is predicted from genomic DNA or cDNA sequence data.

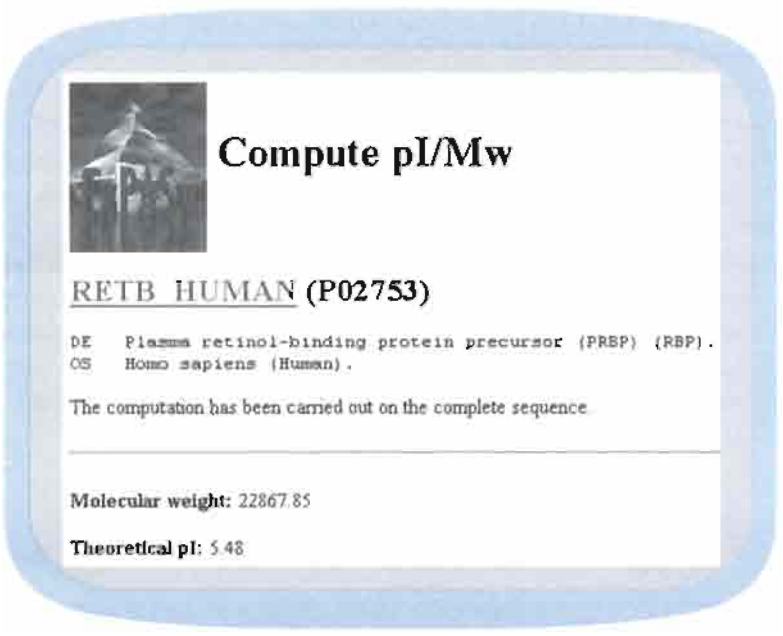


FIGURE 8.10. The Compute pI/Mw server at ExPASy calculates the predicted molecular weight and isoelectric point of input proteins. Here, the values for retinol-binding protein are calculated.

(a)

(b)

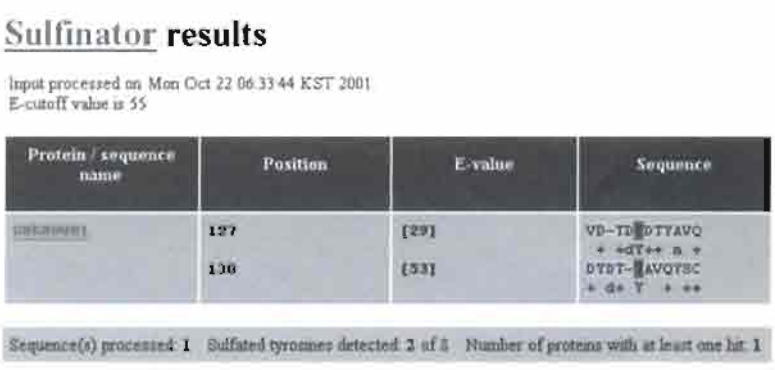


FIGURE 8.11. The ExPASy web server offers a large group of protein analysis tools such as The Sulfinator. (a) A protein sequence is input and (b) the output describes potential sites for sulfation on tyrosine residues. Such information on sulfation, phosphorylation, glycosylation, or other posttranslational modifications may be fundamental in designing experiments to test the function of a protein.

Some proteins with unusual occurrences of particular amino acids are given in Table 8.15 under Web Resources. These proteins may have physical properties (such as pI) that are difficult to predict.

Prediction programs based on primary amino acid sequence data are likely to be more trustworthy than prediction proteins describing secondary protein features such as topology and posttranslational modifications. For proteins with typical amino acid compositions, the prediction of the molecular weight and pI (Fig. 8.10) is likely to be accurate. These protein features can also be experimentally confirmed using techniques such as gel electrophoresis and isoelectric focusing. A prediction algorithm may accurately specify that a protein has a consensus site for phosphorylation or sulfation, but these modifications are not necessarily made in living cells.

What is the accuracy of a program that predicts transmembrane topology? It is easy to use a search tool to find a prediction. However, this is fundamentally a cell biological question, and it requires the tools of cell biology to obtain a clear answer. Many proteins have stretches of 10–25 hydrophobic amino acid residues

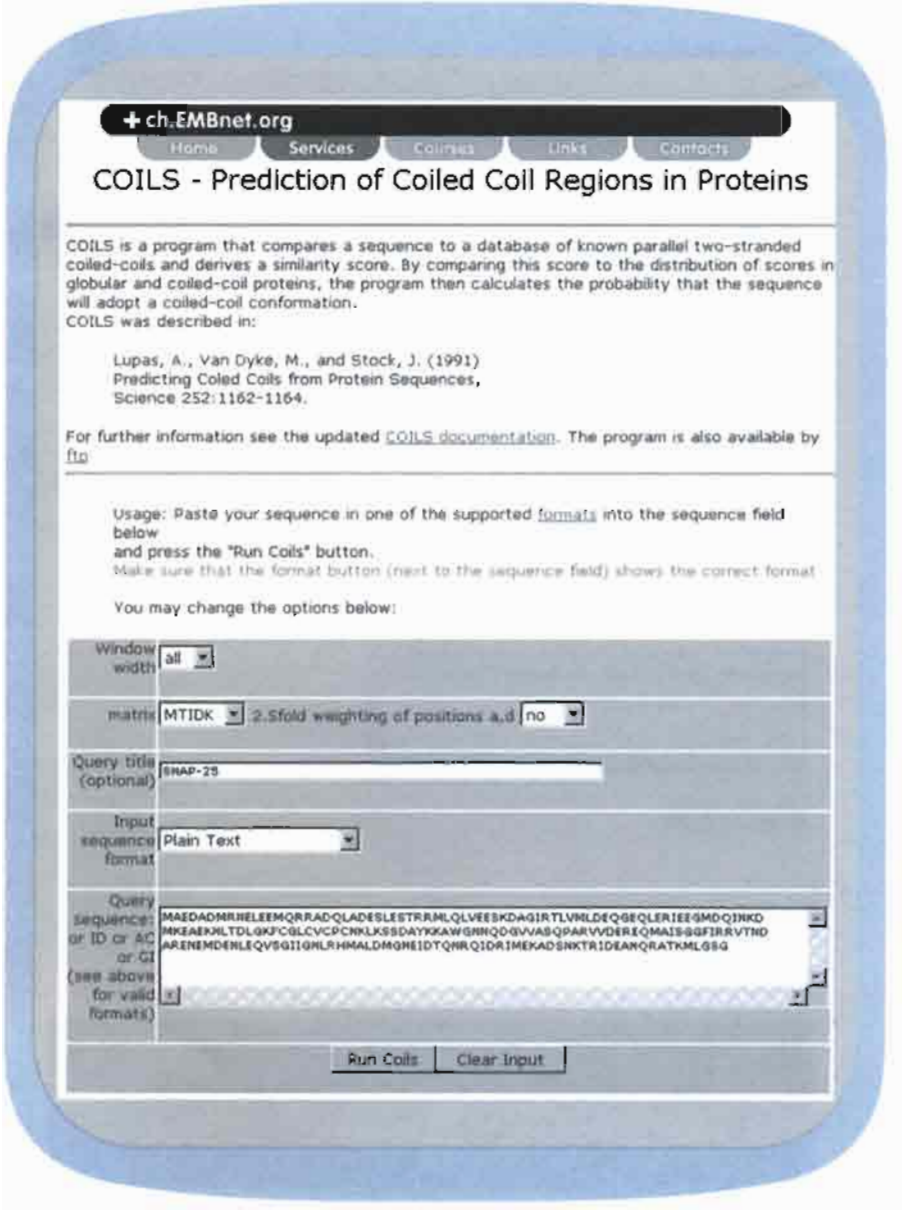


FIGURE 8.12. *The coils program of Lupas et al. (1991) assesses the likelihood that a protein sequence forms a coiled-coil structure.*

that may form transmembrane domains. The most rigorous assessment of the true number of transmembrane domains comes from experimental approaches such as immunocytochemistry. Specific antisera can be raised in rabbits, mice, or other species and used to detect an antigen (such as a cell surface receptor) in a sample affixed to a microscope slide. In unpermeabilized-cells, the antisera can be used to visualize protein regions that are oriented outside the cell. However, when cells are permeabilized with detergent, the antisera can gain access to the cytosol and thus can visualize intracellular (cytoplasmic) domains. Cell biological analyses such as these have been used to experimentally determine the number of transmembrane domains, and in many cases these results contradict the predictions of hydropathy plots (e.g., Ratnam et al., 1986).

Similarly, many programs can predict the existence of glycosylation, phosphorylation, or other sites. These predictions can be extremely valuable in guiding the biologist to make predictions about the possible posttranslational modifications of a protein. These predictions may then be tested experimentally.

We will address the structure of proteins in detail in Chapter 9. Many programs predict secondary-structure features of proteins (see Chapter 9, Table 9.1). One such feature is coiled-coil regions, which are typically associated with protein–protein interaction domains (Lupas et al., 1991; Lupas, 1997) (Figs. 8.12 and 8.13).

Introduction to Perspectives 3 and 4: Gene Ontology Consortium

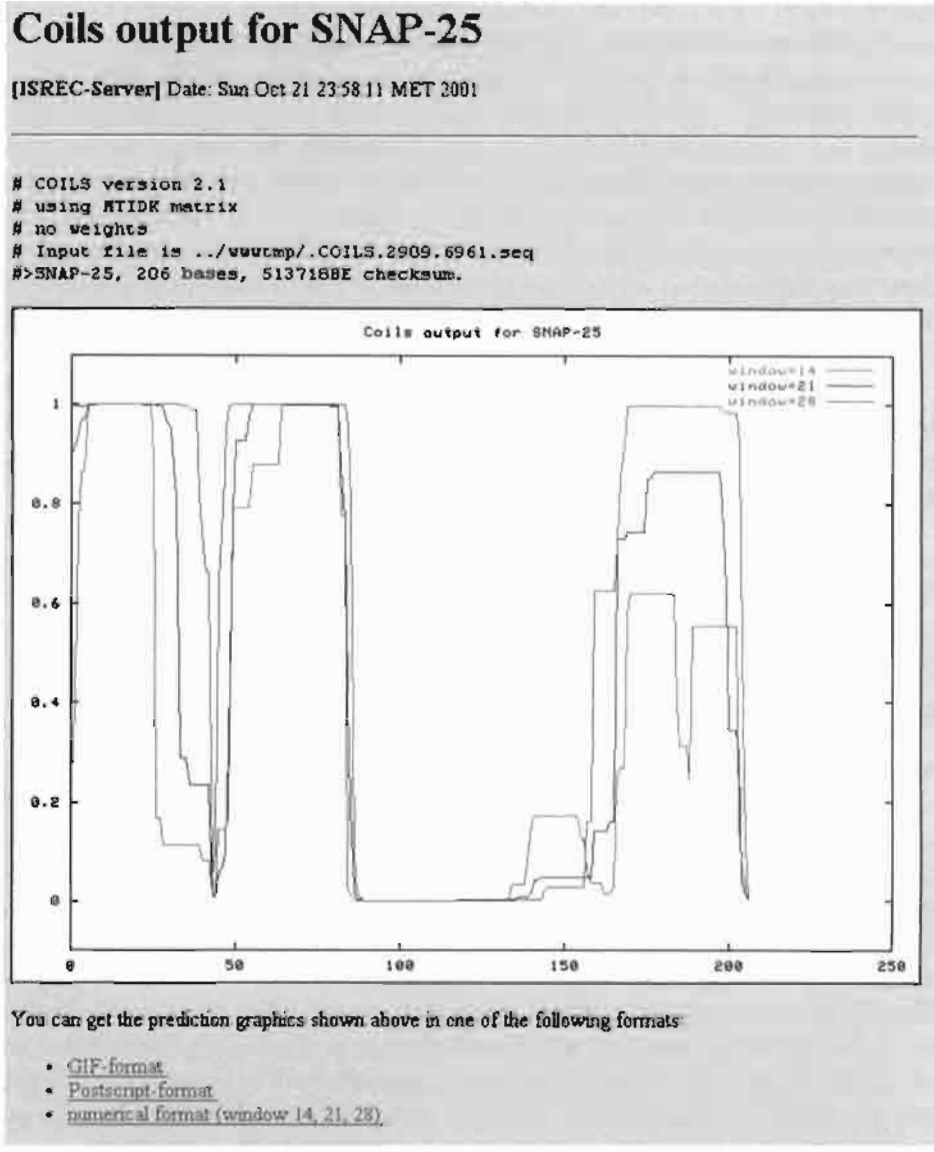
An ontology is a description of concepts. The GO Consortium is a project that compiles a dynamic, controlled vocabulary of terms related to different aspects of genes and gene products (proteins). The consortium was begun by scientists associated with three model organism databases: the *Saccharomyces* Genome Database (SGD), the *Drosophila* genome database (FlyBase), and the Mouse Genome Informatics databases (MGD/GXD) (Ashburner et al., 2000, 2001). Subsequently, databases associated with many other organisms have joined the GO Consortium (Table 8.5). The GO database is not centralized per se but instead relies on external databases (such as a mouse database) in which each gene or gene product is annotated with GO terms. Thus it represents an ongoing, cooperative effort to unify the way genes and gene products are described. There are several web browsers that serve as principal gateways to search GO terms (Table 8.6). Additionally, LocusLink entries at NCBI (Chapter 2) contain GO terms.

The Gene Ontology Consortium main web site is ▸http://www.geneontology.org/.

There are three main organizing principles of GO: (1) molecular function, (2) biological process, and (3) cellular compartment. Molecular function refers to the tasks performed by individual gene products. For example, a protein can be a transcription factor or a carrier protein. Biological process refers to the broad biological goals that a gene product (protein) is associated with, such as mitosis or purine metabolism. Cellular component refers to the subcellular localization of a protein. Examples include nucleus and lysosome. Any protein may participate in more than one molecular function, biological process, and/or cellular component.

Genes and gene products are assigned to GO categories through a process of annotation. The author of each GO annotation supplies an evidence code that indicates the basis for that annotation (Table 8.7). As an example of a GO-annotated protein look at the LocusLink entry for retinol-binding protein (Fig. 8.14*a*). LocusLink entries include a section on function that includes information from

(a)



(b)

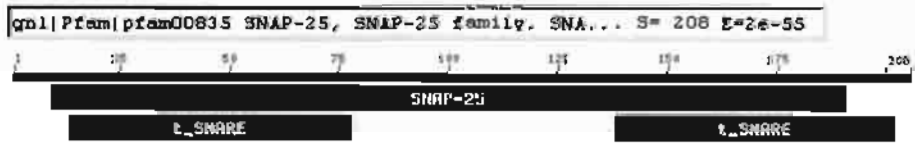


FIGURE 8.13. (a) *Output of the coils program from the ISREC server. The sequence of SNAP-25 (NP_003072) was input. The result depicts the probability that the protein will form a coiled-coil secondary structure motif (y-axis) across the length of the protein (x-axis). Three window sizes are used to estimate the probability of coiled-coil formation. Coiled-coils often represent protein-protein interaction domains. In this case, the coiled-coils of SNAP-25, a peripherally associated plasma membrane protein, allow it to bind tightly to two other proteins (syntaxin and vesicle-associated membrane protein [VAMP/synaptobrevin]) to coordinate synaptic vesicle docking and neurotransmitter release at the presynaptic nerve terminal.* (b) *Upon using SNAP-25 as a query in a blastp search the coiled-coil domains are identified (as so-called t-SNAREs) by the SMART database. This coils site is accessed at* ▸*http://www.isrec.isb-sib.ch/cgi-bin/print_hit_bold.pl/software/COILS_form.html.*

TABLE 8-5 Participating Organizations and Databases in Gene Ontology Consortium

Database or Organization	Organism	Common Name	URL
Berkeley Drosophila Genome Project	*Drosophila melanogaster*	Fly	▶http://www.fruitfly.org/
Compugen			▶http://www.cgen.com/
DictyBase	*Dictyostelium discoideum*	Slime mold	▶http://dictybase.org/
European Bioinformatics Institute (EBI)	Various	–	▶http://www.ebi.ac.uk/
FlyBase	*D. melanogaster*	Fly	▶http://flybase.bio.stanfor.edu/
Genome Knowledge Base (GKB) at Cold Spring Harbor Laboratory	Various	–	▶http://www.genomeknowledge.org/
Gramene	*Oryza sativa;* other grains, monocots	Rice	▶http://www.gramene.org/
Mouse Genome Database (MGD) and Gene Expression Database (GXD)	*Mus musculus*	Mouse	▶http://www.informatics.jax.org/
Pathogen Group at the Wellcome Trust Sanger Institute	Various	–	▶http://www.sanger.ac.uk/Teams/Pathogen/
PomBase	*Schizosaccharomyces cerevisiae*	Fission yeast	▶http://www.sanger.ac.uk/Projects/S_pombe/
Rat Genome Database (RGD)	*Rattus*	Rat	▶http://rgd.mcw.edu/
Saccharomyces Genome Database (SGD)	*Saccharomyces cerevisiae*	Baker's yeast	▶http://genome-www.stanford.edu/Saccharomyces/
The Arabidopsis Information Resource (TAIR)	*Arabidopsis thaliana*	Thale cress	▶http://www.arabidopsis.org/
The Institute for Genomic Research (TIGR)	Various		▶http://www.tigr.org/
WormBase	*Caenorhabditis elegans*	Worm	▶http://www.wormbase.org/

Source: Adapted from ▶http://www.geneontology.org/.

TABLE 8-6 Websites Useful to Access Gene Ontology Data

Browser	Description	URL
AmiGO	GO browser from the Berkeley Drosophila Genome Project	▶http://www.godatabase.org/cgi-bin/go.cgi
Mouse Genome Informatics (MGI) GO Browser	From Jackson Laboratories	▶http://www.informatics.jax.org/userdocs/GO.help.shtml
"QuickGO" at EBI	From the EMBL and European Bioinformatics Institute; integrated with InterPro (Chapter 10)	▶http://www.ebi.ac.uk/ego/manual.html
Expression Profiler (EP) GO Browser	GO browser and analysis tool that is part of the EP suite at the European Bioinformatics Institute	▶http://ep.ebi.ac.uk/EP/
Cancer Gene Anatomy Project (CGAP) GO Browser	From the National Cancer Institute, NIH	▶http://cgap.nci.nih.gov/Genes/AllAboutGO

TABLE 8-7 Evidence Codes for Gene Ontology Project

Abbreviation	Evidence Code	Example(s)
IC	Inferred by curator	A protein is annotated as having the function of a "transcription factor." A curator may then infer that the localization is "nucleus."
IDA	Inferred from direct assay	An enzyme assay (for function); immunofluorescence microscopy (for cellular component)
IEA	Inferred from electronic annotation	Annotations based on "hits" in searches such as BLAST (but without confirmation by a curator; compare ISS)
IEP	Inferred from expression pattern	Transcripts levels (e.g., based on Northern blotting or microarrays) or protein levels (e.g., from Western blots)
IGI	Inferred from genetic interaction	Suppresors; genetic lethals; complementation assays; experiments in which one gene provides information about the function, process, or component of another gene
IMP	Inferred from mutant phenotype	Gene mutation; gene knockout; overexpression; antisense assays
IPI	Inferred from physical interaction	Yeast two-hybrid assays; copurification; co-immunoprecipitation; binding assays
ISS	Inferred from sequence or structural similarity	Sequence similarity; domains; BLAST results that are reviewed for accuracy by a curator
NAS	Nontraceable author statement	Database entries such as a SwissProt record that does not cite a published paper
ND	No biological data available	Corresponds to "unknown" molecular function, biological process, or cellular compartment
TAS	Traceable author statement	Information in a review article or dictionary

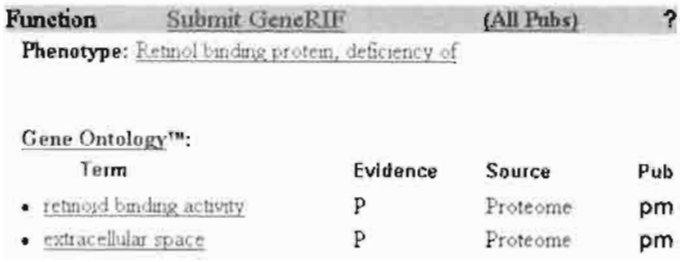
Source: Adapted from ▸http://www.geneontology.org/.

OMIM (Chapter 18), Enzyme Commission nomenclature (see below), and GO terms. For RBP4, the GO terms are retinoid binding (a molecular function) and extracellular space (a cellular component). No entry for biological process is given.

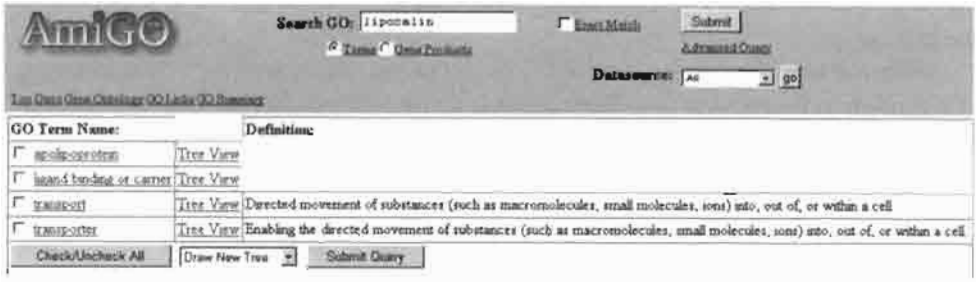
You can also access gene ontology information by entering a query term such as "lipocalin" into a GO web browser (Fig. 8.14*b*). The output includes a graphical tree view (Fig. 8.14*c*). This displays the relationships between the different levels of GO terms. These have the form of a "directed acyclic graph" or network. This differs from a hierarchy in that in a hierarchy each child term can have only one parent, while in a directed acyclic graph it is possible for a child to have more than one parent. A child term may be an instance of its parent term, in which case the graph is labeled "isa," or the child term may be component of the parent term (a "partof" relationship). Examples of these relationships are shown in Fig. 8.14*c*.

We will next consider protein localization and protein function. These topics loosely correspond to the GO categories "cellular component" and "molecular function." Later in the chapter we will discuss protein pathways, although the GO category "biological process" does not refer specifically to pathways.

(a)



(b)



(c)

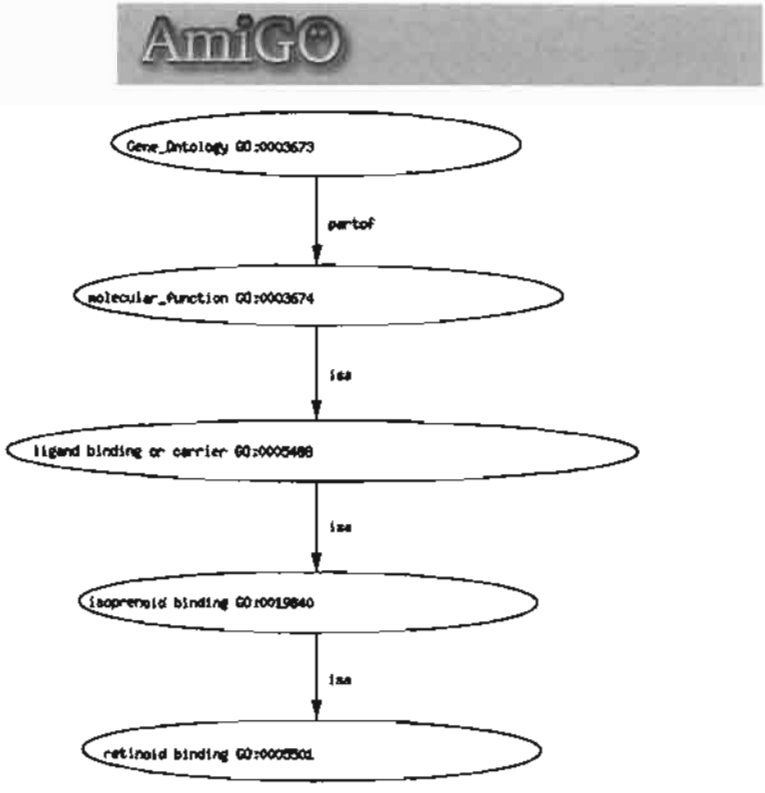


FIGURE 8.14. The GO Consortium provides a dynamic, controlled vocabulary that describes genes and gene products from a variety of organisms. Its three organizing principles are molecular function, biological process, and cellular component. (a) *GO is accessed through a variety of browsers or through LocusLink, as shown for retinol-binding protein.* (b) *Following the link GO terms are shown in the AmiGO browser, including* (c) *a tree view.*

Perspective 3: Protein Localization

The cellular localization of a protein is one of its fundamental properties. Proteins are synthesized on ribosomes from mRNA. Some proteins are synthesized in the cytosol. Other proteins, destined for secretion or insertion in the plasma membrane, are inserted into the endoplasmic reticulum (in eukaryotes) or into the plasma membrane (in prokaryotes). This insertion, which occurs either cotranslationally or posttranslationally, is mediated by the signal recognition particle, an RNA–multiprotein complex (Stroud and Walter, 1999). In the endoplasmic reticulum, proteins may be transported through the secretory pathway to the Golgi apparatus and then to further destinations such as intracellular organelles (e.g., endosomes, lysosomes) or to the cell surface.

In eukaryotic cells, the intracellular organelles account for up to 95% of the cell's membranes.

Proteins may further be secreted into the extracellular milieu. The trafficking of a protein to its appropriate destination is achieved by transport in secretory vesicles. These vesicles are typically 75–100 nanometers in diameter, and they transport soluble or membrane-bound cargo to specific compartments.

We may also distinguish two main categories of proteins based upon their relationships to phospholipid bilayers: those that are soluble and exist in the cytoplasm, in the lumen of an organelle, or in the extracellular environment and those that are membrane attached, associated with a lipid bilayer. Those proteins associated with membranes may be integral membrane proteins (having a span of 10–25 hydrophobic amino acid residues that traverse the lipid bilayer) or they may be peripherally associated with membranes (attached via a variety of anchors; see below).

Many proteins defy categorization into one static location in the cell. For example, the annexins and the low-molecular-weight GTP-binding proteins are families of proteins that migrate between the cytosol and a membrane compartment. This movement typically depends on the presence of dynamically regulated cellular signals such as calcium.

Proteins are often targeted to their appropriate cellular location because of intrinsic signals embedded in their primary amino acid sequence. For example, the sequence KDEL (lysine–aspartic acid–glutamate–leucine) at the carboxy terminus of a soluble protein specifies that it is selectively retained in the endoplasmic reticulum. Other targeting motifs have been identified for import into mitochondria

Results of Subprograms

```
PSG:  a new signal peptide prediction method
      N-region:  length 2;  pos.chg 1;  neg.chg 0
      H-region:  length 14;  peak value  10.03
      PSG score:   5.63

GvH:  von Heijne's method for signal seq. recognition
      GvH score (threshold: -2.1):   3.93
      possible cleavage site: between 16 and 17

>>> Seems to have a cleavable signal peptide (1 to 16)
```

FIGURE 8.15. *The PSORT server (http://psort.nibb.ac.jp/) provides a web-based query form to predict the subcellular location of a protein. The program searches for sorting signals that are characteristic of proteins localized to particular compartments. The output of a search using retinol-binding protein shows that there is strong evidence for a signal peptide with a cleavage site between amino acid residues 16 and 17. Such a signal peptide characterizes proteins that enter the secretory pathway where some (such as RBP) are secreted outside the cell.*

or peroxisomes and for endocytosis. However, these motifs are not as invariant as KDEL.

Several web-based programs predict the intracellular localization of any individual protein sequence (see Web Resources, Table 8.16). For example, PSORT accurately predicts the signal sequence at the amino terminus of retinol-binding protein (Fig. 8.15). This signal peptide is characteristic of proteins that enter the secretory pathway in the endoplasmic reticulum. A more general problem is identifying a transmembrane region. Programs such as TMPred (Web resources, Table 8.17) accomplish this (Fig. 8.16). As described above (pp. 236–237), algorithms that describe protein features such as membrane topology must be used with caution.

Michael Snyder and colleagues attempted the first proteome-scale analysis of protein localization (Kumar et al., 2002). They cloned cDNA encoding several thousand proteins from the budding yeast *S. cerevisiae*, incorporating epitope tags into the carboxy or amino termini. This directed cloning approach allowed them to systematically evaluate the location of many specific proteins of interest. As a complementary strategy, they used random transposon-mediated tagging of genes.

Kumar et al. (2002) generated over 13,000 yeast strains and determined the subcellular localization of 2744 yeast proteins by immunofluorescence microscopy using monoclonal antibodies directed against the epitope tags. Many proteins of unknown function were assigned to intracellular locations. For example, if a protein is localized to the yeast peroxisome, then this suggests it may function in fatty acid metabolism.

An epitope tag is a short protein fragment, such as the nine-amino-acid hemagglutinin (HA) peptide, that is attached covalently to a protein of interest. An antiserum that detects the epitope tag can then be used to localize the protein of interest by immunofluoresence microscopy or to purify the protein (and its binding partners) by immunoprecipitation with an anti-epitope tag antiserum.

Bacterial transposons are mobile DNA elements that can be randomly inserted into genomic DNA. The transposons can be modified to incorporate an epitope tag. This strategy is practically simpler than directed cloning of specific yeast cDNAs, although from an experimental point of view transposon-tagged proteins are difficult to localize in cells.

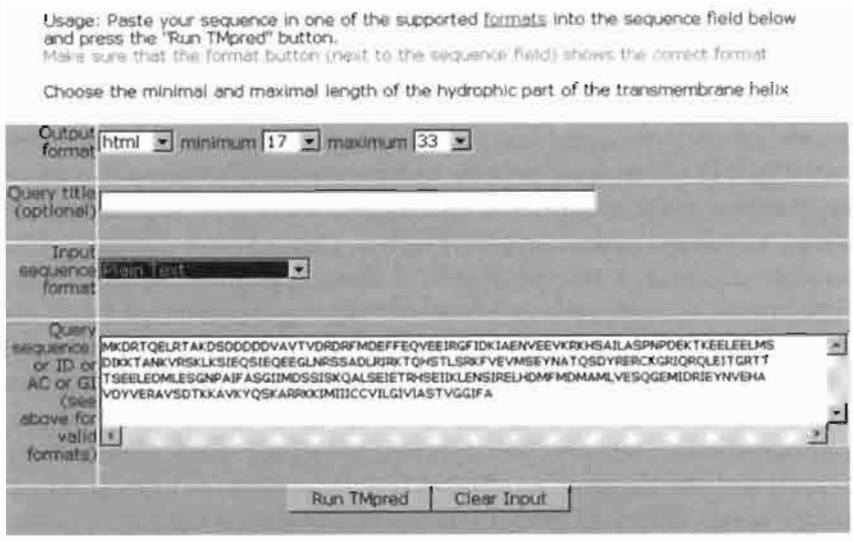
Perspective 4: Protein Function

We have described bioinformatics tools to describe protein families, their physical properties, and the cellular localization of proteins. A fourth aspect of proteins is their function. Function is defined as the role of a protein in a cell (Jacq, 2001). Each protein is a gene product that interacts with the cellular environment in some way to promote the cell's growth and function. We can consider the concept of function from several perspectives (Fig. 8.17):

A database with 2900 fluorescence micrographs of the Kumar et al. (2002) data is available at ▸http://ygac.med.yale.edu.

- A protein has a biochemical function synonymous with its molecular function. For an enzyme, the biochemical function is to catalyze the conversion of one or more substrates to product(s). For a structural protein such as actin or tubulin, the biochemical function is to influence the shape of a cell. For a transport protein, the biochemical function is to carry a ligand from one location to another. (Such a transport role may even occur in the absence of a requirement for an energy source such as ATP—in such a way, retinol-binding protein transports retinol through serum.) For a hypothetical protein that is predicted to be encoded by a gene, the biochemical function is unknown but is presumed to exist. There are thought to be no proteins that exist without a biochemical function.
- Functional assignment is often made based upon homology. If a hypothetical protein is homologous to an enzyme, it is often provisionally assigned that enzymatic function. This type of functional assignment is best viewed as a hypothesis that must be tested experimentally.
- Function may be assigned based upon structure (Chapter 9). If a protein has a three-dimensional fold that is related to that of a protein with a known

(*a*)



(*b*)

```
2 possible models considered, only significant TM-segments used

-----> slightly prefered model: N-terminus inside
 1 strong transmembrane helices, total score : 3757
 # from   to length score orientation
 1  266  284 (19)    2757 i-o

------> alternative model
 1 strong transmembrane helices, total score : 2690
 # from   to length score orientation
 1  266  288 (23)    2690 o-i
```

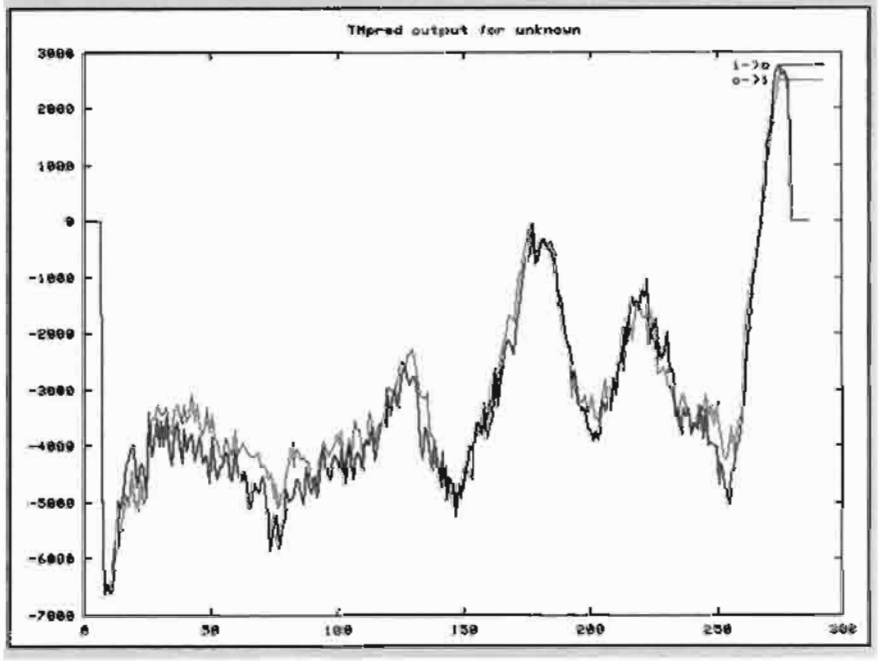


FIGURE 8.16. A variety of programs can predict transmembrane spans of proteins. (a) *The sequence of syntaxin (NP.004594) was input into the TMpred program (http://www.ch.embnet.org/software/TMPRED_form.html).* (b) *The output shows the model (which has been experimentally verified) showing a single transmembrane domain at the carboxy terminus of the protein.*

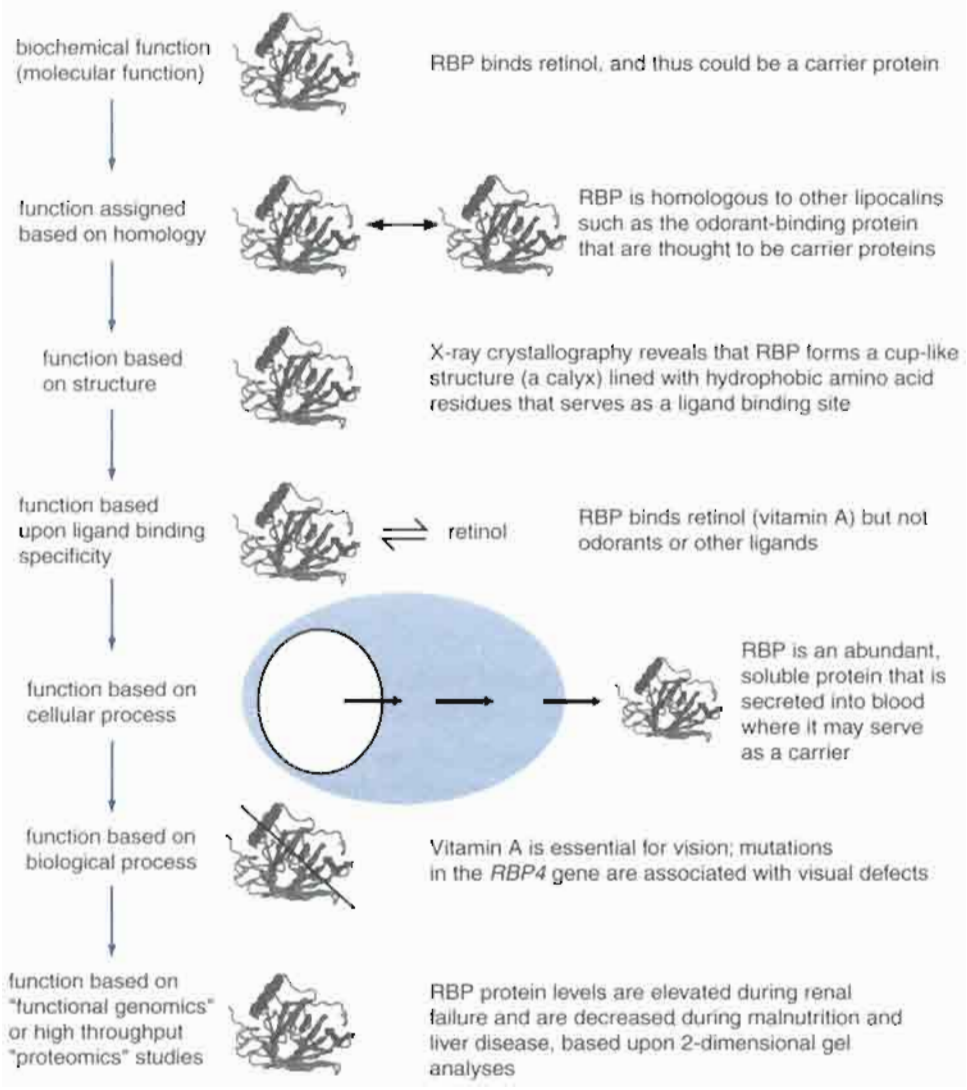


FIGURE 8.17. Protein function may be analyzed from several perspectives. Retinol-binding protein (RBP) is used as an example.

function, this may be the basis for functional assignment. Note, however, that structural similarity does not necessarily imply homology, and homology does not necessarily imply functional equivalence.

- All proteins function in the context of other proteins and molecules. Thus a definition of a protein's function may include its ligand (if the protein is a receptor), its substrate (if the protein is an enzyme), its lipid partner (if the protein interacts with membrane), or any other molecule with which it interacts. The odorant-binding protein (OBP) is a lipocalin that binds a variety of odorants in nasal mucus, suggesting that the binding properties of the protein are central to its function (Pevsner et al., 1990). However, the biological function of OBP is not known from its ligand-binding properties alone. The protein could transport odorants toward the olfactory epithelium to promote sensory perception, it could carry odorants from the olfactory epithelium to facilitate odorant clearance, or it could metabolize odorants.
- Many proteins function as part of a distinct biochemical pathway such as the Krebs cycle, in which discrete steps allow the cell to perform a complex

TABLE 8-8 Functional Assignment of Proteins Based upon Their Enzymatic Activity: Partial List of EC Classification System (Release 27.0, October 2001)

EC Number	Description of Class	Number of Enzymes	Example of Subclass
1. -. -. -	Oxidoreductases	1003	
1. 1. -. -	–	–	Acting on the CH–OH group of donors
1. 2. -. -	–	–	Acting on the aldehyde or oxo group of donors
2. -. -. -	Transferases	1076	
2. 1. -. -	–	–	Transferring one-carbon groups
3. -. -. -	Hydrolases	1125	
4. -. -. -	Lyases	356	
5. -. -. -	Isomerases	156	
6. -. -. -	Ligases	126	

Source: From http://kr.expasy.org/cgi-bin/enzyme-search-cl.

task. Other examples are fatty acid oxidation in peroxisomes or proteolytic degradation that is accomplished by the proteasome.

- Proteins function as part of some broad cell biological process. Cells divide, grow, and senesce; neurons have axons that display outgrowth, pathfinding, target recognition, and synapse formation; and all cells secrete molecules through discrete pathways. All cellular processes require proteins in order to function, and each individual protein can be defined in the context of the broad cellular function it serves.
- Protein function can be considered in the context of all the proteins that are encoded by a genome—that is, in terms of the proteome. The term *functional genomics* refers to the attempt to use experimental approaches and/or computational tools to analyze the role of many hundreds or thousands of expressed genes (i.e., RNA transcripts). Since the ultimate product of transcription is a protein, the term functional genomics is sometimes applied to large-scale studies of protein function.

The Gene Ontology Consortium (Ashburner et al., 2000, p. 27) defines a biological process as "a biological objective to which the gene or gene product contributes. A process is accomplished via one or more ordered assemblies of molecular functions. Processes often involve a chemical or physical transformation, in the sense that something goes into a process and something different comes out of it."

Thus protein function can be defined in many ways. Many proteins are enzymes. The Enzyme Commission (EC) system provides a standardized nomenclature for almost 4000 enzymes (Table 8.8). When a genome is sequenced and a potential protein-coding sequence is identified, homology of that protein to an enzyme with a defined EC listing provides a specific, testable hypothesis about the biochemical function of that hypothetical protein.

Another broader approach to the functional assignment of proteins is provided by the Clusters of Orthologous Groups (COG) database (Chapter 14). The functional groups defined by this system are listed in Table 8.9. While the COG database has initially focused on prokaryotic genomes, the general categories are relevant to basic cellular processes in all living organisms. Many other functions that are unique to eukaryotes, such as apoptosis and complex developmental processes, are not represented in the COG scheme.

Apoptosis is programmed cell death. It occurs in a variety of multicellular organisms, both as a normal process in development and as a homeostatic mechanism in adult tissues. Apoptosis can be triggered by external stimuli (such as infectious agents or toxins) or by internal agents such as those causing oxidative stress.

TABLE 8-9 Functional Classification of Proteins in Clusters of Orthologous Groups Database

General Category	Function	Clusters of Orthologous Groups	Domains
Information storage and processing	Translation, ribosomal structure, and biogenesis	217	6,449
	Transcription	133	5,442
	DNA replication, recombination, and repair	184	5,337
Cellular processes	Cell division and chromosome partitioning	32	842
	Posttranslational modification, protein turnover, chaperones	109	3,155
	Cell envelope biogenesis, outer membrane	155	4,079
	Cell motility and secretion	133	3,110
	Inorganic ion transport and metabolism	160	5,112
	Signal transduction mechanisms	96	3,623
Metabolism	Energy production and conversion	223	5,584
	Carbohydrate transport and metabolism	170	5,257
	Amino acid transport and metabolism	233	8,383
	Nucleotide transport and metabolism	85	2,364
	Coenzyme metabolism	154	4,057
	Lipid metabolism	75	2,609
	Secondary-metabolite biosynthesi, transport, and catabolism	62	2,754
Poorly characterized	General function prediction only	449	11,948
	Function unknown	752	6,431

Source: From ▸http://www.ncbi.nlm.nih.gov/COG/, September, 2002.

PROTEOMICS: BIOINFORMATIC TOOLS FOR HIGH-THROUGHPUT PROTEIN ANALYSIS

Classical biochemical approaches to protein function involve an assay for the function of a protein (such as its enzymatic activity or a bioassay for its influence on a cellular process). This assay may be used as the basis of a purification scheme in which the protein is purified to homogeneity. Thousands of proteins have been studied individually with this approach. Each protein has its own personality in terms of biochemical properties and its propensity to interact with a variety of resins that separate proteins on the basis of size, charge, or hydrophobicity.

In addition to the study of individual proteins, high-throughput analyses of thousands of proteins are possible (Molloy and Witzmann, 2002). We will describe four such approaches: two-dimensional gel electrophoresis, affinity chromatography with mass spectrometry, the yeast two-hybrid system, and a computational approach called "Rosetta Stone" that is based upon the analysis of genomic DNA sequences. These approaches are presented in Figure 8.18.

The functions of most proteins are unknown. Even for relatively well studied model organisms such as *Escherichia coli* and *S. cerevisiae,* functions have been assigned to only about half of all proteins. The high-throughput proteomics projects attempt to assign function on a large scale, identifying the presence of proteins in particular physiological conditions or identifying protein–protein interaction partners.

There are additional high-throughput approaches to proteomics. Protein microarrays, analogous to DNA microarrays, consist of affinity reagents (such as specific antibodies) that are attached to a solid support (MacBeath, 2002). Such technology has not yet reached widespread use because of the inherent difficulty in maintaining the structure (and function) of immobilized proteins. Tissue microarrays represent another high-throughput approach that is particularly well suited to molecular pathology studies (Kononen et al., 1998; Kallioniemi et al., 2001). A tissue microarray typically consists of several hundred (or thousand) tissue specimens immobilized on a slide in an orderly array. These samples can be probed in parallel to detect and quantify DNA, RNA, or protein targets.

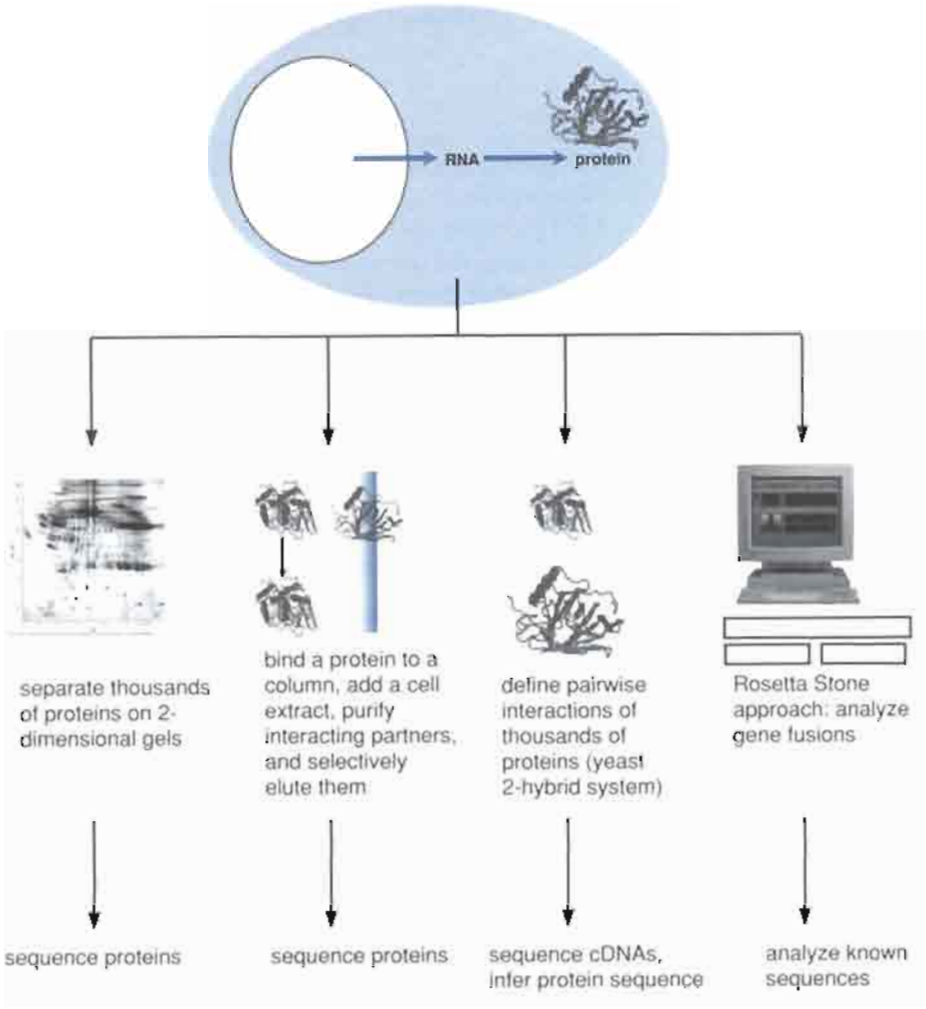


FIGURE 8.18. Approaches to high-throughput protein analysis. Additional strategies such as protein chips (high-density arrays containing immobilized proteins) are also under development.

The most comprehensive approaches to understanding gene and protein function in a single organism have come from studies of the budding yeast *Saccharomyces cerevisiae*. We will discuss these approaches in detail in Chapter 15.

Two-Dimensional Gels Electrophoresis

Polyacrylamide gel electrophoresis (PAGE) is a premier tool for the analysis of protein molecular weight. Proteins (like nucleic acids) possess a charge and thus migrate when introduced into an electric field. Proteins are denatured and electrophoresed through a matrix of acrylamide that is inert (so it does not interact with the protein) and porous (so that proteins can move through it). The velocity of a protein as it migrates through an acrylamide gel is inversely proportional to its size, and thus a complex mixture of proteins can be separated in a single experiment. Proteins are almost always electrophoresed through acrylamide under denaturing conditions in the presence of the detergent sodium dodecyl sulfate (SDS), so this technique is commonly abbreviated SDS-PAGE.

O'Farrell (1975) greatly extended the capabilities of this technology by combining it with an initial separation of proteins based on their charge. In the first step, proteins are separated by isoelectric focusing. A gel matrix (or strip) is produced that contains ampholytes spanning a continuous range of pH values, usually between pH 3 and 11. Each protein is zwitterionic, and when electrophoresed, it migrates to the position at which its total net charge is zero. This is the isoelectric point (abbreviated pI) at which the protein stops migrating. A complex mixture of

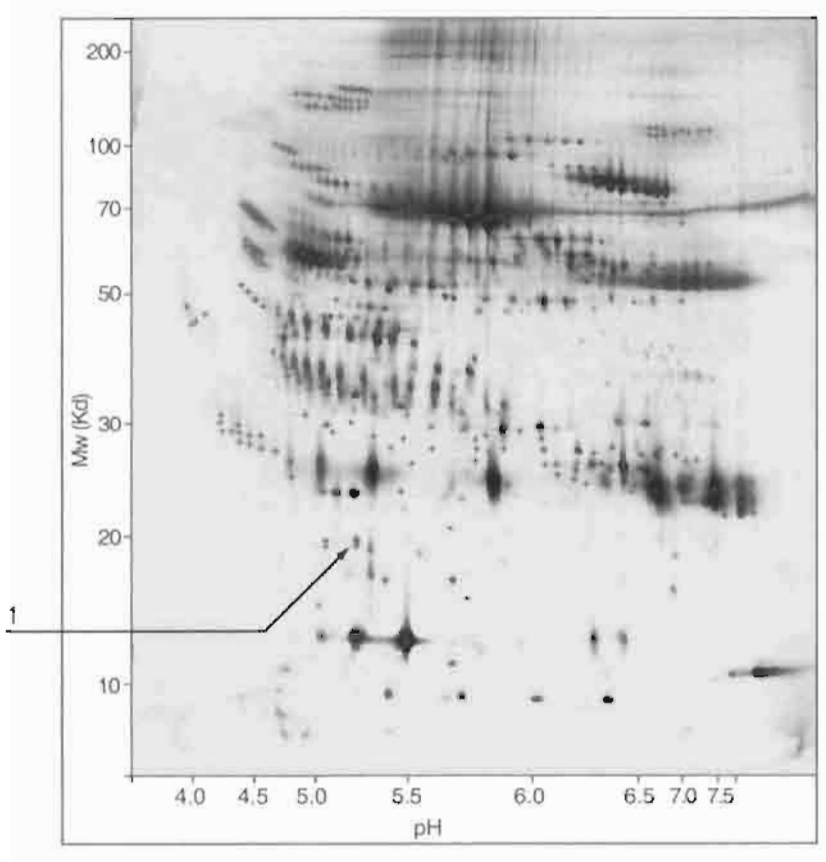


FIGURE 8.19. A two-dimensional gel profile of cerebrospinal fluid from ExPASy. Each of the crosses is linked to a description of a protein that has been identified by microsequencing. Retinol-binding protein is indicated (arrow 1), having an isoelectric point of about 5.2 and a mass of 20 kDa.

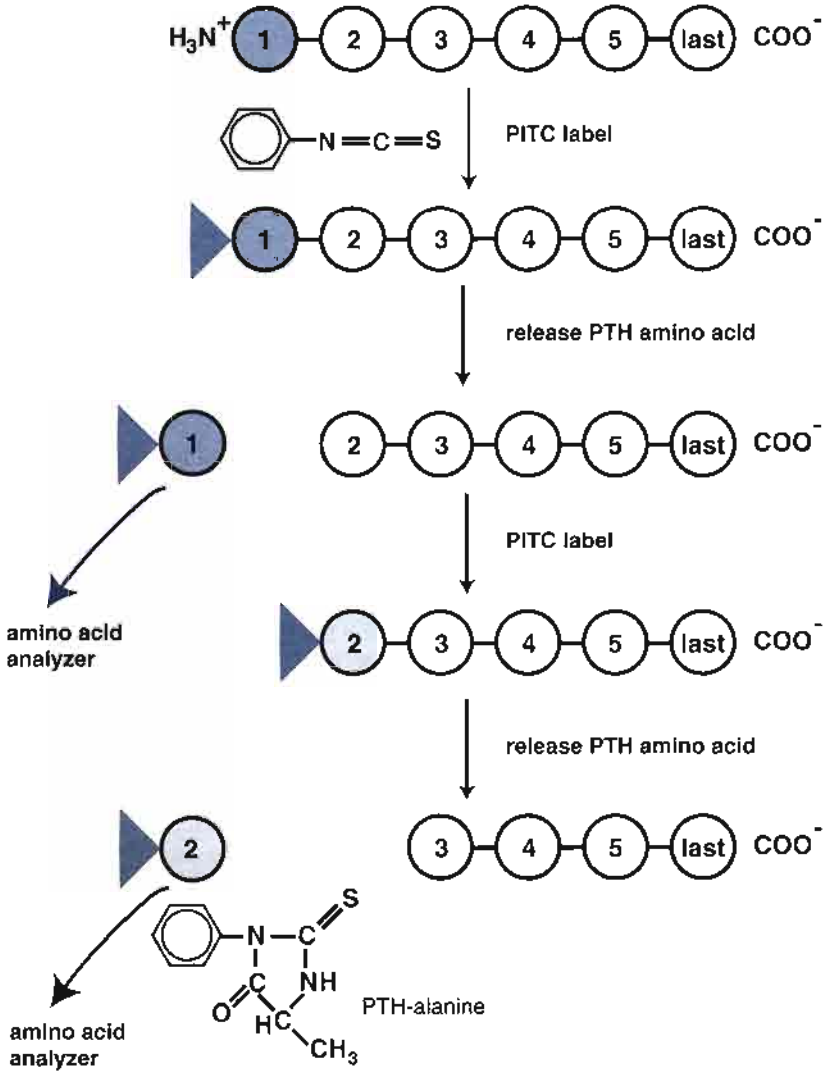
proteins may thus be separated based upon charge, and this corresponds to the first dimension of two-dimensional gel electrophoresis. In the second dimension, proteins are separated by SDS–PAGE. The technique of two-dimensional gel electrophoresis has matured into an important technology used to analyze proteomes (Dunn, 2000).

An example of a two-dimensional gel profile is shown in Figure 8.19. Several hundred micrograms of protein from human cerebrospinal fluid were separated by pI (on the x axis) by isoelectric focusing, then by size (on the y axis) by SDS–PAGE. Thousands of proteins may be visualized with a protein-binding dye such as silver nitrate or Coomassie blue. Note that several proteins are especially abundant. Many proteins have a characteristic pattern of spots that spread along the first dimension. This is a "charge train" that usually represents a series of variants of a protein with differing amounts of charged groups such as phosphates that are covalently attached.

A key property of two-dimensional protein gels is that the individual proteins may be identified by direct protein microsequencing (Box 8.1) or by sensitive mass spectroscopy techniques such as matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) spectroscopy (Farmer and Caprioli, 1998) (Box 8.2). Applications of two-dimensional SDS–PAGE include a description of hundreds of proteins in human and rat brain (Langen et al., 1999; Fountoulakis et al., 1999) and an analysis of aberrant protein expression profiles in bladder tumors (Østergaard et al.,

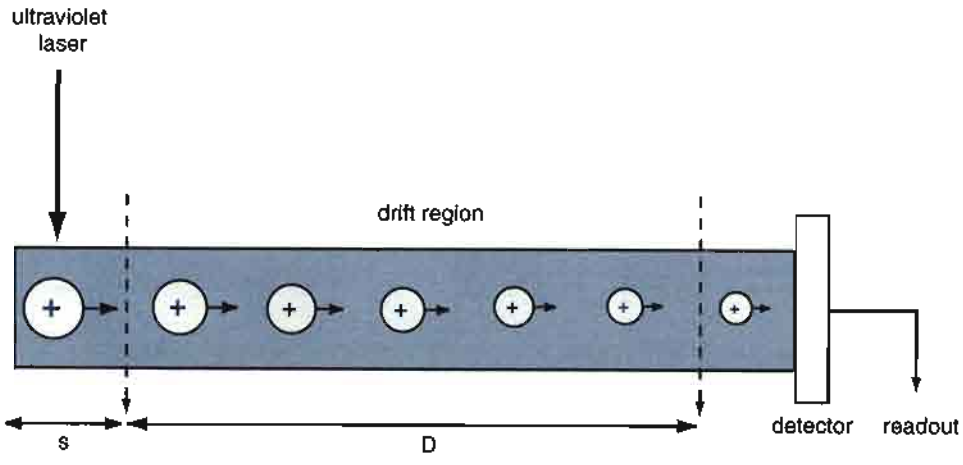
BOX 8-1
Protein Sequencing by Edman Degradation

One obtains a portion of the amino acid sequence of a protein by transferring it to a specialized membrane polyvinylidene fluoride or (PVDF), then submitting it to a core facility for microsequencing by sequential Edman degradations. Eighty percent of the time, the amino terminus is blocked (e.g., acylated and unavailable for Edman degradations). A standard procedure is to proteolyze (e.g., trypsinize) the protein, purify the proteolytic fragments by reverse-phase high-performance liquid chromatography (HPLC), confirm the purity and mass of the fragments by mass spectrometry (MALDI-TOF spectroscopy), and perform Edman degradations. The Edman process is illustrated for a protein fragment of six amino acids. The first amino acid reacts through its amino terminus with phenylisothiocyanate (PITC). This forms a residue derivitized with phenylthiohydantoin (PTH) that is identified in an amino acid analyzer. The cycle is repeated with successive amino-terminal amino acids. The structure of PTH-alanine is shown as an example. The typical result is a readout of 10–20 amino acids. The corresponding protein and gene can be evaluated by performing BLAST searches (Chapter 4).



BOX 8-2
Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Spectroscopy MALDI-TOF

Spectroscopy is a technique to measure the mass of protein samples and other macromolecules. A sample is placed in a matrix of material that absorbs ultraviolet light. A laser is fired at the sample in the source region (*s*), and in the context of the matrix the sample becomes ionized. Some of the protein samples evaporate (i.e., desorption occurs). The ionization occurs in the presence of an electric field that accelerates the ions into a long drift region (*D*). The acceleration of each protein fragment is proportional to the mass of the ion.



1997). Grünenfelder et al. (2001) analyzed protein synthesis during the cell cycle of the bacterium *Caulobacter crescentus* and detected about 25% (979) of all the predicted gene products. Many of these were degraded during a single cell cycle. It will be of great interest to correlate large-scale protein findings with mRNA expression data obtained from microarrays.

We addressed the issue of mRNA-protein correlations in Chapter 6.

ExPASy provides the main database for information on two-dimensional gel electrophoresis (Hoogland et al., 1999; Sanchez et al., 2001). Information is available for gels from a variety of organisms and experimental conditions. These profiles may be queried by choosing a two-dimensional gel map (Fig. 8.20) which links to an image such as that shown in Figure 8.19. This database is also searchable by other criteria such as keyword. Upon entering "retinol" as a search term, four gels in which retinol-binding protein has been identified are shown (Fig. 8.21). By further selecting any of these two-dimensional gels, an image is produced highlighting the protein of interest (Fig. 8.22).

The two-dimensional gel (2DG) approach has several limitations.

- It is not amenable to high-throughput processing of many samples in parallel.
- Only the most abundant proteins in a sample are usually detected. Hydrophobic proteins, including proteins with transmembrane domains, are underrepresented on two-dimensional gels.
- It requires considerable expertise to reliably generate consistent results. In comparing two 2DG profiles, if the polyacrylamide gels vary even slightly in composition or if the samples are electrophoresed under differing conditions,

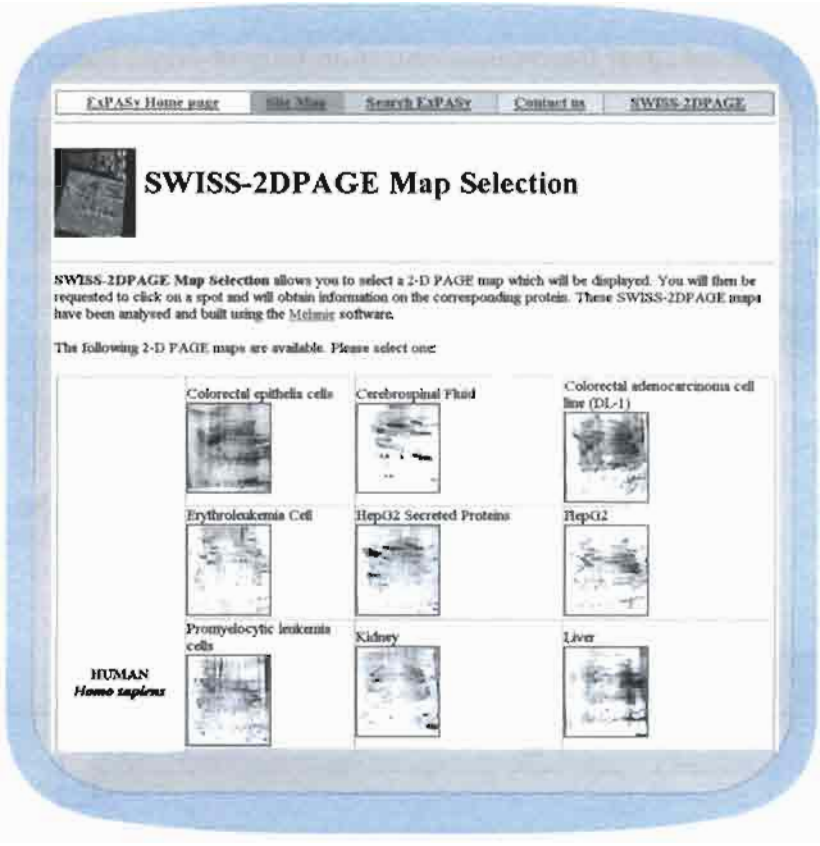


FIGURE 8.20. The Swiss-2DPAGE (two-dimensional polyacrylamide gel electrophoresis) website at ExPASy offers images of two-dimensional protein gels from a variety of organisms, including human (shown above), mouse, plants, and bacteria. Each gel profile can be expanded and explored in depth by clicking on it.

it can be difficult to accurately align the protein spots. An important technical advance in the reproducibility of 2DG electrophoresis was the introduction of immobilized pH gradients, preformed on dry strips, that replaced an older system of pH gradient formation with ampholytes.

Affinity Chromatography and Mass Spectrometry

Affinity chromatography is a technique in which a ligand such as a protein is chemically immobilized to a matrix on a column. Two groups employed a strategy of identifying thousands of multiprotein complexes in the yeast *S. cerevisiae* (Gavin et al., 2002; Ho et al., 2002; reviewed in Kumar and Snyder, 2002). Each group selected large numbers of "bait" proteins containing a tag that allowed each bait to be introduced into yeast, where they could form native protein complexes. After complexes were allowed to form under physiologically relevant conditions, the bait was extracted, copurifying associated proteins. These protein complexes were resolved by one-dimensional SDS-PAGE. Thousands of individual protein gel bands (from experiments with many different bait proteins) were excised from the gel with a razor, digested with trypsin to form relatively small protein fragments, and identified by MALDI-TOF mass spectrometry (Box 8.2).

Employing this strategy, Gavin et al. (2002) obtained 1167 yeast strains expressing tagged proteins, from which they purified 589 tagged proteins and identified 232 protein complexes. Ho et al. (2002) selected 725 bait proteins and also detected thousands of protein–protein associations. In each case, a large number of the

The Gavin et al. (2002) data are available at ▸http://yeast.cellzome.com. The Ho et al. (2002) data are available at ▸http://mdsp.com/yeast, and their data have been deposited in BIND (Biomolecular Interaction Network Database) at ▸http://www.bind.ca/. For additional protein interaction databases, see Table 8.18 under Web Resources.

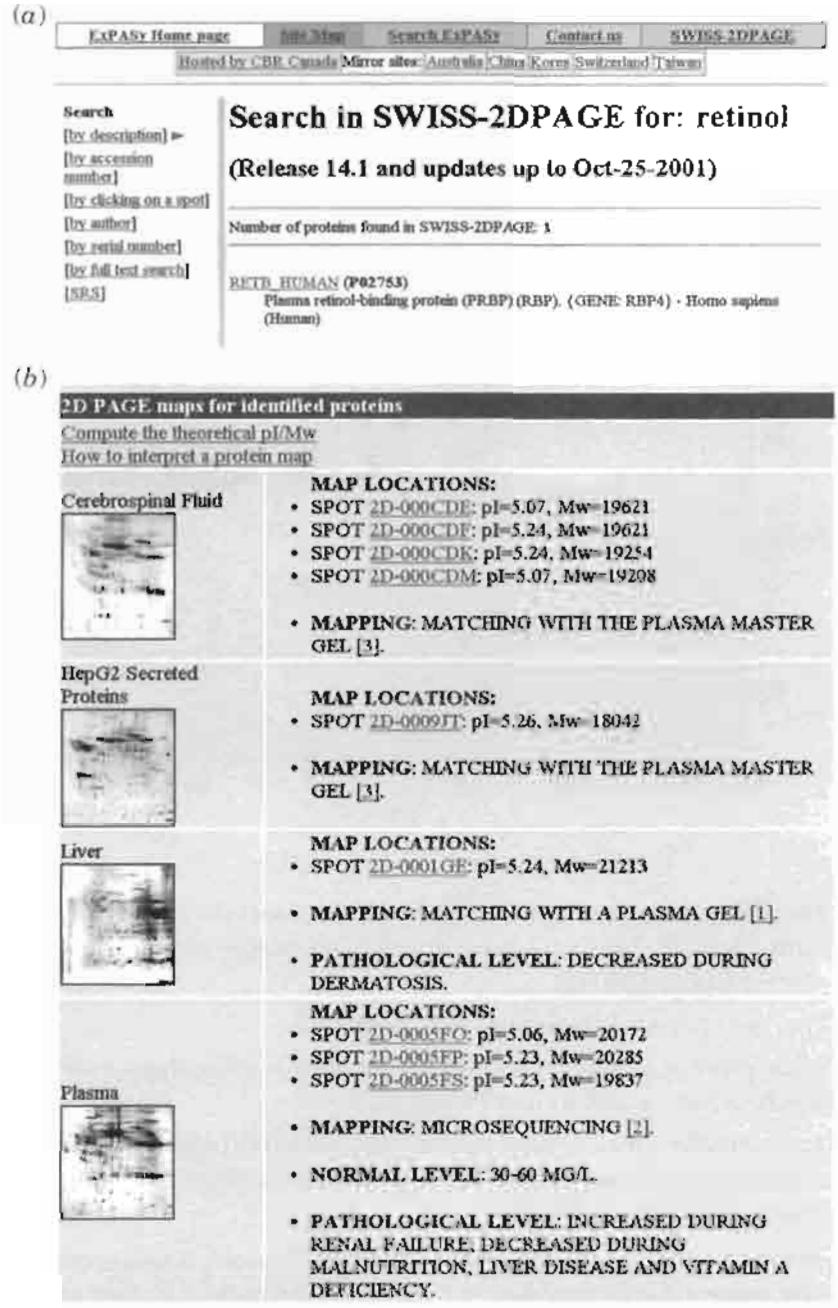


FIGURE 8.21. *The SWISS-2DPAGE website is searchable by keyword.* (a) *A search with the word "retinol" yields a result that* (b) *links to a series of two-dimensional protein gels known to contain RBP. Note that additional information is provided; in plasma, elevated RBP levels are associated with several pathological conditions. Note also that in cerebrospinal fluid RBP has been identified in four locations with slightly varying molecular weight and isoelectric focusing values. This heterogeneity typically reflects different run conditions or slight differences in the posttranslational modifications added to a protein, such as differing amounts of phosphorylation.*

protein complexes that were identified included proteins of previously unknown function, highlighting the strength of these large-scale approaches.

This experimental strategy entails a number of assumptions, including reasons for false-positive results (biologically nonsignificant interactions) and false-negative results (missed biological interactions). False-negative results may occur for the following reasons:

- The bait that is introduced into yeast cells must be localized properly. Failure of the bait to behave in its native condition could explain why some previously known interactions were not observed.

Issues of false-positive and false-negative results have been discussed by Schächter (2002) in the context of yeast two-hybrid screens (see below).

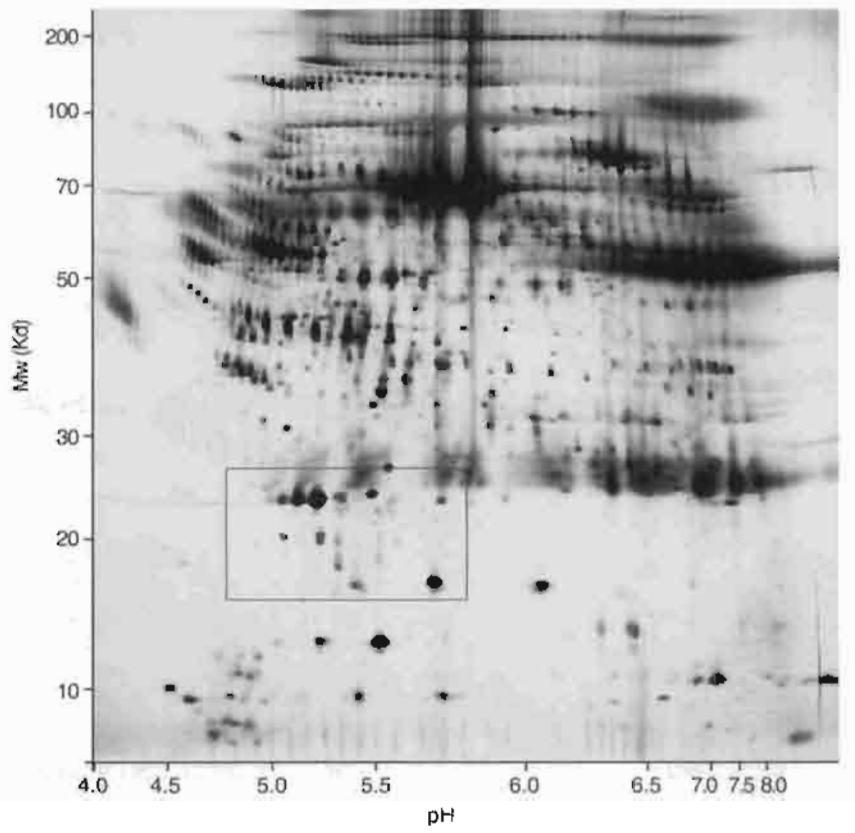


FIGURE 8.22. The location of RBP is indicated on a two-dimensional protein gel derived from human cerebrospinal fluid (CSF). This image was obtained by clicking on the CSF image (Fig. 8.21b). Note also the differences between this image of CSF proteins and the one shown in Fig. 8.19. Such differences routinely occur because of differences in the conditions of separating proteins by isoelectric focusing and/or gel electrophoresis (e.g., time of electrophoresis, buffer concentrations, ampholyte concentrations, percentage of acrylamide employed, protein-staining method). Nonetheless the two profiles in Figure 8.19 and here are closely similar.

- The affinity tag must not interfere with the function of the bait protein. In some cases, the tags are large (e.g., 20 kDa) relative to the average size of a protein (about 50 kDa).
- Transient protein interactions may be missed.
- Some protein complexes require highly specific physiological conditions in which to form and thus may be missed.
- There may be a bias against hydrophobic proteins (which are more difficult to purify than soluble proteins) and low-molecular-weight proteins below 15 kDa (Gavin et al., 2002).

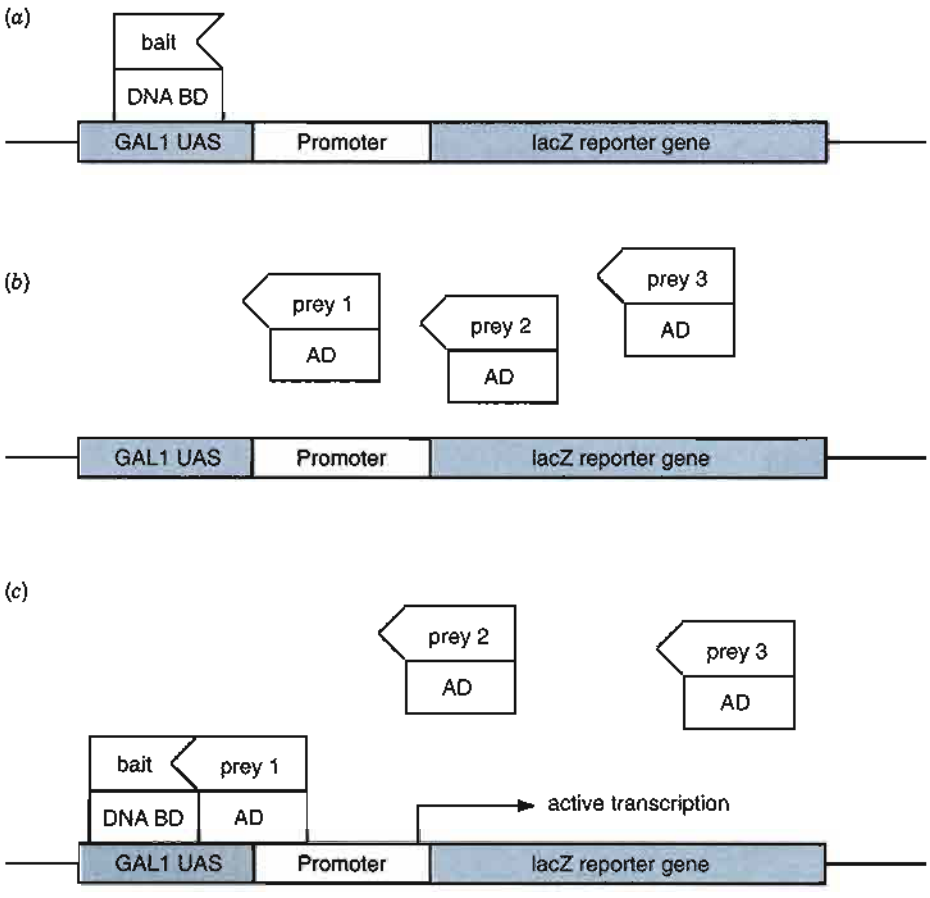
False-positive results may also occur for a variety of reasons. Some proteins may be inherently susceptible to nonspecific binding interactions (i.e., they are "sticky"). Proteins that are denatured may bind nonspecifically.

Yeast Two-Hybrid System

The yeast two-hybrid system is a method used to identify protein–protein interactions (Fields and Song, 1989). The assay is extremely versatile and has been used to identify protein-binding partners in many species. It is based upon the fact that the yeast *GAL4* transcriptional activator is composed of two independent activation and binding domains (see Box 8.3). The cDNA encoding a protein of interest (the "bait") is fused to the *GAL4* DNA binding domain. A large collection of cDNAs (a library consisting of various "prey") is cloned into a vector containing the *GAL4* activation domain. Alone, the *GAL4* DNA binding domain does not activate transcription.

BOX 8-3
Yeast Two-Hybrid System

The yeast two-hybrid system allows the identification of the binding partners of a protein. A cDNA encoding a protein of interest (such as huntingtin, the protein that is mutated in Huntington's disease) is used as a "bait" to identify interacting proteins in a library of cDNAs encoding human proteins expressed in brain ("prey"). A construct containing huntingtin cDNA, fused to a DNA binding domain (BD), is introduced into yeast cells. The BD interacts with a yeast *GAL1* upstream activating sequence (UAS), but in the absence of an appropriate activator domain (AD) a *lacZ* reporter gene is not activated [see (*a*) below]. A library of thousands of cDNAs is created, each fused to an activation sequence, but these alone are also unable to activate a reporter gene [see (*b*)]. When a clone from the library (AD fused to prey 1) binds to the bait/DNA BD construct, the activator domain is able to activate transcription of the *lacZ* reporter gene. This reporter allows identification of plasmid DNA from these yeast cells, and the prey 1 cDNA is sequenced. There may be many different binding partners identified from a yeast two-hybrid library. In one application of this technology, Li et al. (1995) identified huntingtin-associated protein (HAP-1), a protein enriched in brain that may affect the selective neuropathology of expanded polyglutamine repeats in Huntington's disease (Chapter 18).



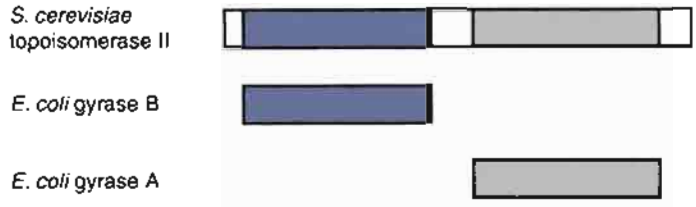


FIGURE 8.23. The Rosetta Stone method has been used to predict functional interactions between proteins based upon analysis of genomic DNA sequences. Genome sequences are scanned for the presence of independent genes from one organism (e.g. E. coli *gyrases A and B) that occur in orthologs as part of a single open reading frame. The presence of such a fusion event is interpreted as evidence that the two proteins are part of a functionally related pathway. For the organism that has fused the two genes, an evolutionary advantage could be conferred by the usefulness of co-regulating expression of the two functional units, and/or there may be entropic benefit from the presence of high local concentrations of both proteins.*

But when the bait binds to another fusion protein expressed from the cDNA library, the proximity of the two proteins enables transcription of a *GAL4* reporter gene. The name "two-hybrid" system refers to the use of two recombinant proteins that must interact.

In addition to the strategy of using a bait protein to screen a library, the yeast two-hybrid system has been used to measure the interaction of a known bait protein with individual, cloned prey proteins. In this way a set of many protein-protein interactions can be assayed. Compared to screening libraries, this approach has the advantage of systematically testing a matrix of possible protein-protein interactions, while it has the disadvantage of not allowing the discovery of novel interacting partners that might be found in a complex cDNA library.

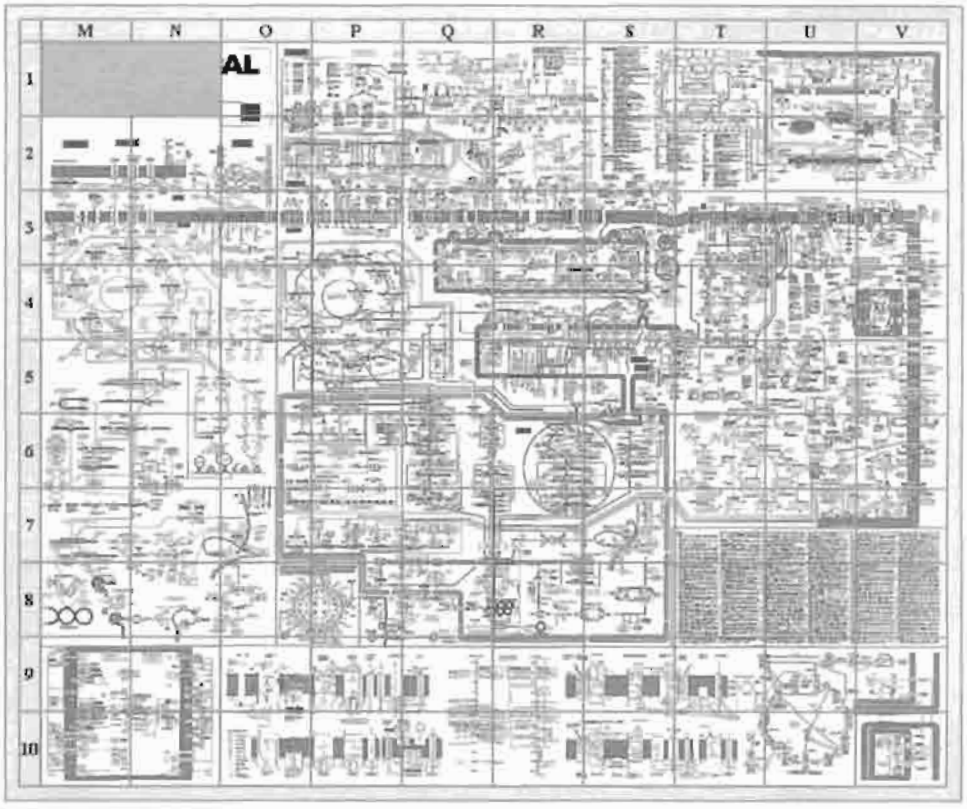
Yeast two-hybrid system technology has been applied to comprehensive analyses of essentially all possible pairwise protein-protein interactions in the yeast *S. cerevisiae*. Uetz et al. (2000) described 957 interactions involving 1004 yeast proteins, while Ito et al. (2001) identified 4549 interactions among 3278 proteins. These data sets are useful to define possible pathways of interacting proteins. Surprisingly, only about 20% of these two data sets overlap. The lack of concordance between these data sets may be due to differences in the physiological conditions in the studies, or to different sources of false positive and false negative errors (discussed below).

The Ito et al. (2001) data set is available at ▸ http://genome.c.kanazawa-u.ac.jp/Y2H. In another proteomewide application of two-hybrid technology, Rain et al. (2001) generated a large interaction map for *Helicobacter pylori* proteins.

A major difference between the yeast two-hybrid strategy and the affinity chromatography approach is that the yeast two-hybrid system is only used to detect pairwise interactions between proteins. In contrast, an affinity chromatography approach allows subunits consisting of many proteins to be isolated and identified.

As with any high-throughput screening technology, there are many issues concerning false-positive and false-negative results (i.e., reliability and coverage). In addition to the issues described for comprehensive affinity purification studies (see pp. 255–256), false-negative results may occur if the protein-protein interactions fail to occur in the specialized environment of the yeast nucleus. False-positive results may occur in a yeast two-hybrid screen when a bait protein autoactivates a reporter gene. Some sticky prey proteins may nonspecifically bind to and activate many bait proteins. Careful analysis of two-hybrid results allows these sources of false-positive and false-negative results to be reduced, for example by identifying promiscuous binding proteins.

(a)



(b)

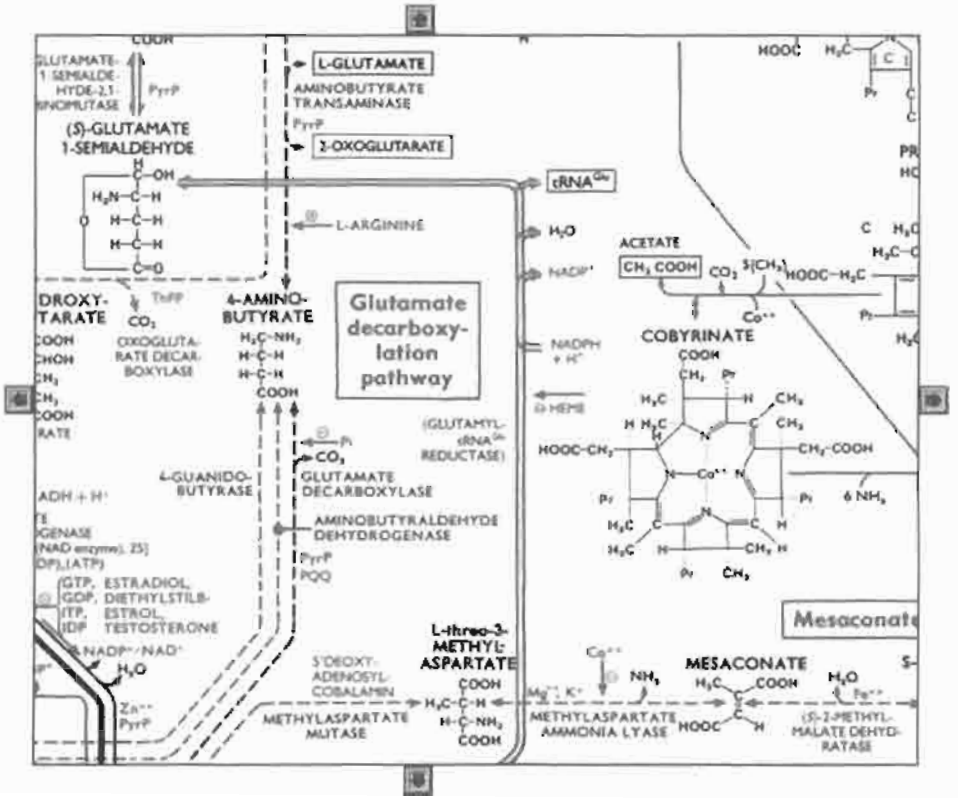


FIGURE 8.24. (a) *The ExPASy website includes an online version of biochemical pathway maps from Boehringer Mannheim.* (b) *A search for glutamate (the major excitatory amino acid neurotransmitter in brain) results in a detailed map. The full map is available from* ▸ *http://biochem.boehringer-mannheim.com/prodinfo.fst.htm?/techserv/metmap.htm.*

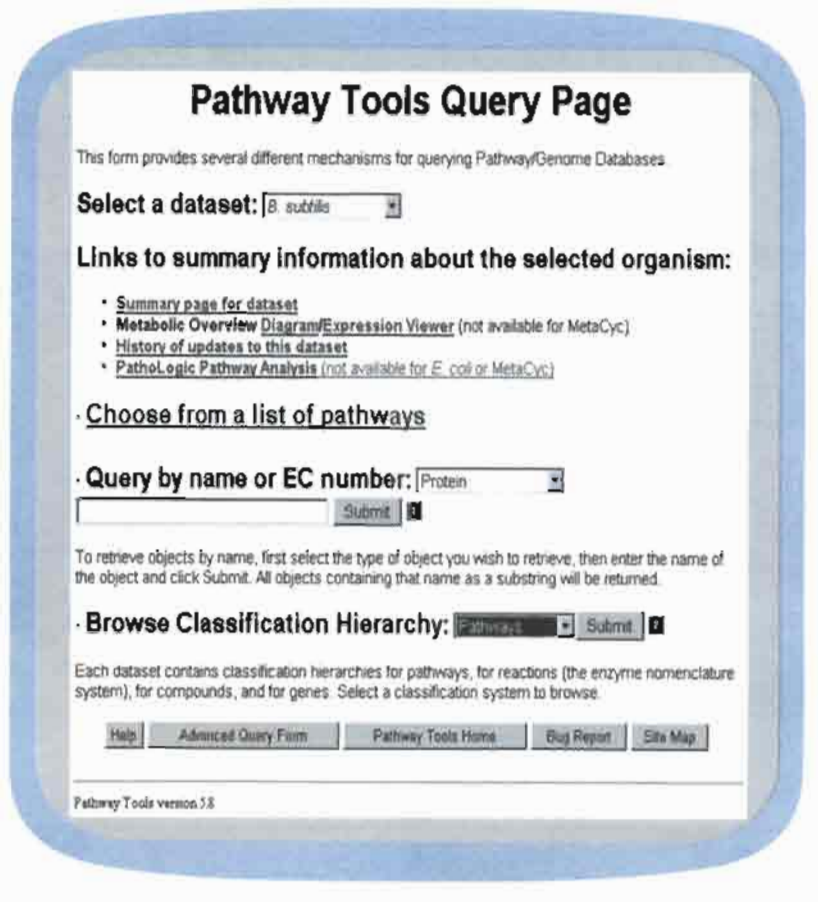


*FIGURE 8.25. The EcoCyc/MetaCyc query page allows access to these databases of pathways. The dataset allows investigation of several organisms (*Bacillus subtilis, Chlamydia trachomatis, E. coli, Helicobacter pylori, Haemophilus influenzae, Mycobacterium tuberculosis, Mycoplasma pneumoniae, Pseudomonas aeruginosa, S. cerevisiae, Treponema pallidum*).*

Rosetta Stone Approach

Several groups have adopted a computational approach to protein function prediction. Marcotte and colleagues (1999a,b) as well as Enright et al. (1999) hypothesized that some pairs of interacting proteins are encoded by two distinct genes in one genome that have fused into a single gene in another genome (Fig. 8.23). Marcotte et al. scanned multiple genomes and identified 6809 such cases in *E. coli* and 45,502 in *S. cerevisiae.* This domain function analysis has been called the Rosetta Stone approach. The Rosetta Stone approach makes the prediction that protein pairs generated from gene fusions have related biological functions. For example, they may function in the same protein complex, pathway, or biological process. The approach also predicts possible protein-protein interactions. An organism that has fused two genes encoding biologically related proteins may benefit from an entropic contribution afforded by increased effective concentrations of the two proteins in a local environment. It is also possible that domain fusion does not involve functionally related proteins but occurs for other reasons.

Rosetta Stone refers to the ancient tablet that contains a "fusion" of three distinct scripts (hieroglyphic, demotic Greek, and classical Greek); the hieroglyphics could be deciphered once all three were discovered together.

Bioinformatic Approaches to Cellular Pathways

Information about the roles of many proteins in a cell can be integrated in databases and visualized with protein network maps (Schächter, 2002). A pathway is a linked set of biochemical reactions (Karp, 2001). The motivation behind making pathway maps is to visualize complex biological processes.

Summary: The primary data source for this dataset was the full genome sequence of Saccharomyces cerevisiae [1]. This dataset was derived computationally using the PathoLogic program [2], with some additional manual curation. Manual curation has not occurred since 1999. Note that each YeastCyc gene contains a WWW link to the SGD database. Please refer to SGD for more detailed information about yeast gene products.

Genetic Element	Total Genes	Protein Genes	RNA Genes	Unidentified ORFs	Size (bp)
Chromosome I	111	107	4	54	230,209
Chromosome II	437	423	14	205	813,138
Chromosome III	225	214	11	130	315,341
Chromosome IV	826	796	30	409	1,531,974
Chromosome V	304	280	24	133	576,870
Chromosome VI	143	133	10	74	270,148
Chromosome VII	599	560	39	287	1,090,936
Chromosome VIII	290	279	11	141	562,638
Chromosome IX	233	223	10	122	439,885
Chromosome X	414	386	28	201	745,440
Chromosome XI	350	333	17	161	666,448
Chromosome XII	571	536	35	306	1,078,172
Chromosome XIII	507	486	21	256	924,430
Chromosome XIV	429	415	14	222	784,330
Chromosome XV	582	562	20	297	1,091,283
Chromosome XVI	505	487	18	267	948,061
Mitochondrial Chromosome	0	0	0	0	78,520
Total:	6526	6220	306	3265	12,147,823

Pathways:	87
Enzymatic Reactions:	600
Transport Reactions:	1
Polypeptides:	6241
Protein Complexes:	42
Enzymes:	509
Transporters:	3
Compounds:	457
Transcription Units:	0
tRNAs:	275

FIGURE 8.26. *The EcoCyc database includes summary information for each organism. A portion of the result for* S. cerevisiae *is presented here.*

What data are appropriate for pathway maps? Proteomewide definitions of protein–protein interactions may be obtained from experimental data (e.g., from the results of yeast two-hybrid screens) or ab initio (e.g., using the Rosetta Stone approach). For some screens (including the yeast two-hybrid system), information may be accumulated about the interactions of particular domain(s) within a protein that are responsible for interactions in addition to information about the interactions of full-length proteins.

Several web servers provide pathway maps (Table 8.18 under Web Resources). The Boehringer Mannheim company produces a comprehensive chart of biochemical pathways that is available on-line at ExPASy (Fig. 8.24).

The EcoCyc/MetaCyc databases provide pathway maps for a variety of organisms. EcoCyc is a model organism database for *E. coli* strain K12 (Karp et al., 2002a)

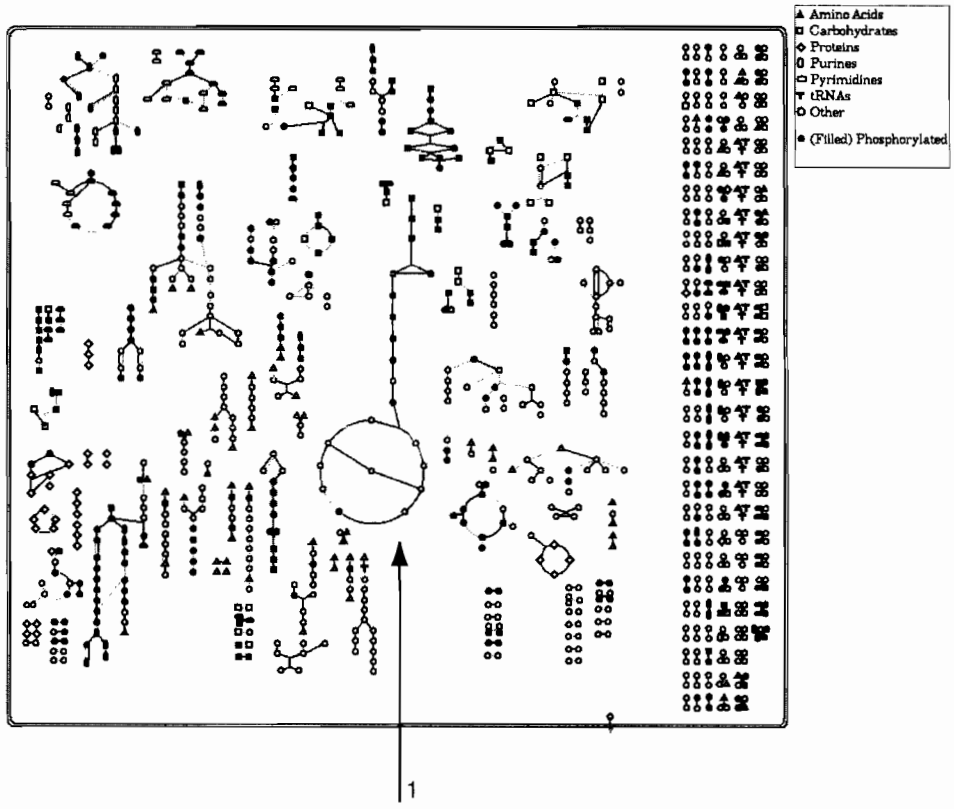


FIGURE 8.27. *The EcoCyc database allows the visualization of entire metabolic pathways. Here,* S. cerevisiae *metabolism is displayed. The Krebs cycle is indicated (arrow 1).*

(Chapter 14). MetaCyc describes about 450 pathways and 1100 enzymes occurring in 158 organisms (Karp et al., 2002b). EcoCyc/MetaCyc includes a variety of query options (Fig. 8.25) and summary data for each organism (Fig. 8.26). A particular pathway may be examined, showing information on genes, their encoded protein products, as well as metabolites and other molecules involved in the pathway (Figs. 8.27–8.29).

You can access the EcoCyc pathway database at ▸http://ecocyc.org. The URL for MetaCyc is ▸http://ecocyc.org/ecocyc/metacyc.html.

Protein interaction networks are often extrapolated between species. Information from a study in yeast is used to infer information about the corresponding network that is presumed to exist in another organism, such as mouse or human. As an example of this approach, Paley and Karp (2002) used several *E. coli* metabolic pathway databases in EcoCyc and MetaCyc to predict the corresponding metabolic pathways in *H. pylori.* They found substantial overlap between pathways predicted by algorithms and those predicted by expert, manual analysis of pathways. Each approach results in both false-positive and false-negative predictions. One of the basic assumptions of pathway comparisons between species is that orthologous proteins can be reliably identified.

The *H. pylori* pathway database is available at ▸http://ecocyc.org.

Another pathway database is offered by the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000). The KEGG database includes a gene database with annotations of orthologs between various species and a pathway database with hundreds of metabolic (and other) pathway maps.

KEGG is available at ▸http://www.genome.ad.jp/kegg/. Release 23 (July 2002) includes about 350,000 genes from 14 eukaryotes (including human, mouse, and rat) and 88 bacteria and archaea.

The KEGG pathway database contains hundreds of metabolic and regulatory pathway maps. The main categories are described in Table 8.10. Selecting one of these, human neurodegenerative disorders, one can find a pathway description of amyotrophic lateral sclerosis (ALS; Lou Gehrig's disease) (Fig. 8.30*a*). Mutations

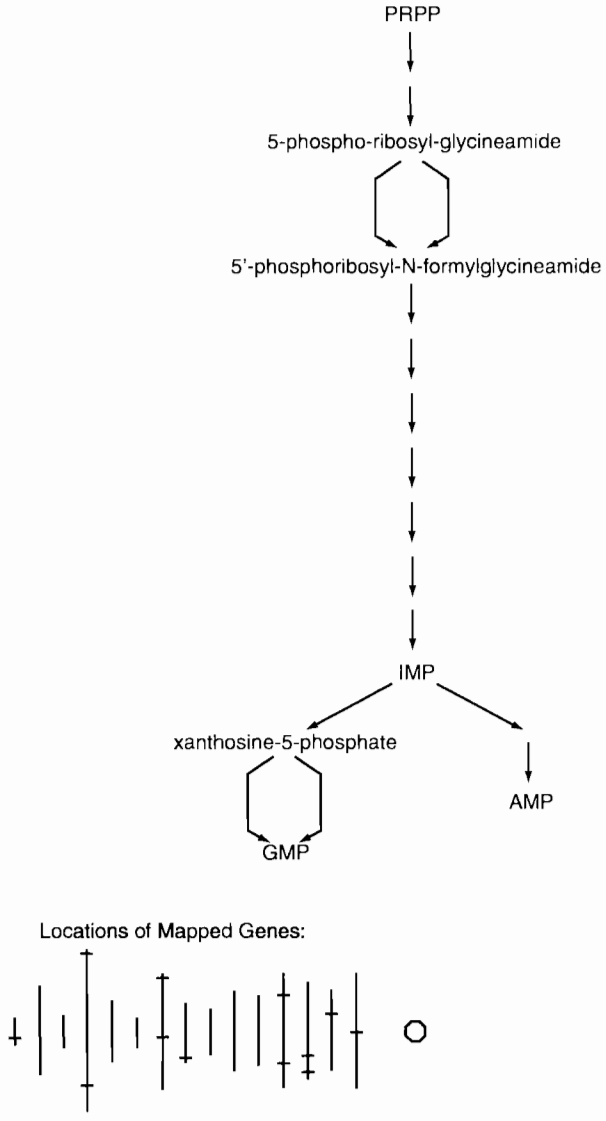
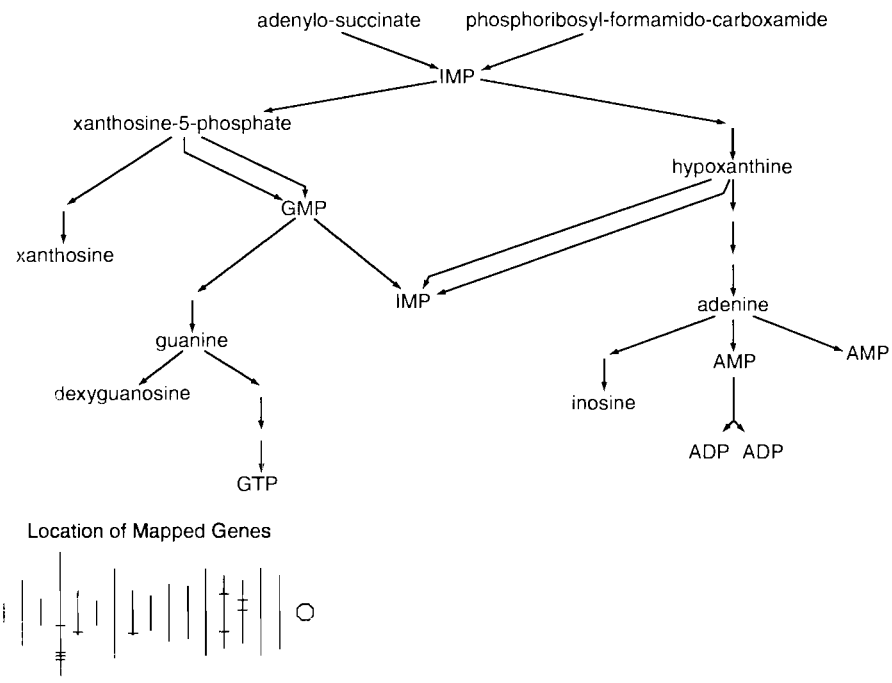


FIGURE 8.28. *Phosphoribosyl pyrophosphate pathway of the EcoCyc database.*

in the superoxide dismutase gene, *SOD1*, are a common cause of this debilitating disease. *SOD1* is an enzyme that normally converts the toxic oxygen metabolite superoxide ($O_2^{-\cdot}$) into hydrogen peroxide and water. As shown in the KEGG pathway map, *SOD1* has been shown to interact directly and indirectly with a variety of other proteins, such as those involved in apoptosis (programmed cell death). Clicking on *SOD1,* one finds an entry describing the protein and nucleotide sequence as well as several external links such as the Enzyme Commission number (Fig. 8.30*b*). A further link to this *SOD1* entry shows additional information, including the results of protein prediction tools for transmembrane domains, protein-sorting signals, and motifs (Fig. 8.30*c*).

This example of *SOD1* highlights a strength of KEGG: Its coverage of a broad range of proteins and cellular processes is comprehensive. The example also serves to show that some processes described in KEGG are likely to be organism specific. KEGG is based primarily on data generated from bacterial genomes, and pathways

(*a*)

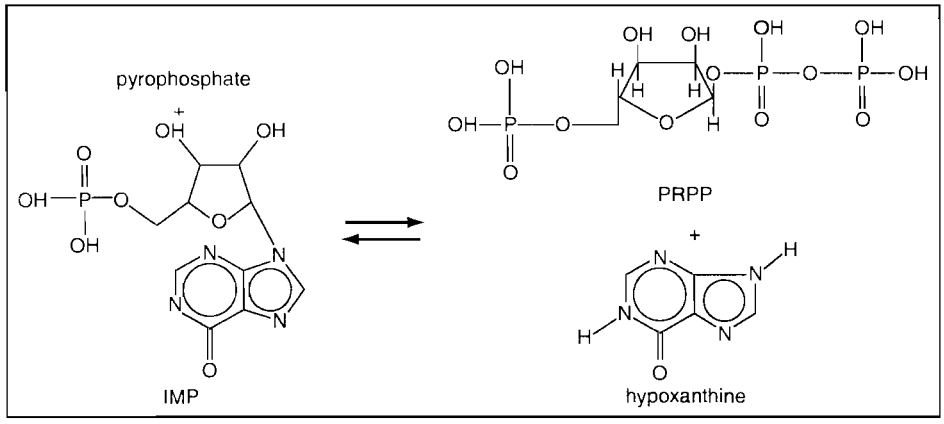


(*b*) **S. *cerevisiae* Reaction: 2.4.2.8**

Superclasses: 2.4.2 -- PENTOSYLTRANSERASES

Hypoxanthine Phosphoribosyltransferase : HPT1

In pathway: nucleotide metabolism



Gene-Reaction Schematic:

Unification Links: ENZYME:2.4.2.8

FIGURE 8.29. (a) *The EcoCyc database includes individual pathways.* (b) *By clicking on individual components of the pathway, a detailed view of the substrates and products is shown.*

TABLE 8-10 Main Categories of Metabolic and Regulatory Pathways in KEGG Database

Hundreds of pathway maps are available within these categories.

Metabolic pathways
Carbohydrate metabolism
Energy metabolism
Lipid metabolism
Nucleotide metabolism
Amino acid metabolism
Metabolism of other amino acids
Metabolism of complex carbohydrates
Metabolism of complex lipids
Metabolism of cofactors and vitamins
Biosynthesis of secondary metabolites
Biodegradation of xenobiotics
Regulatory pathways: genetic information processing
Transcription
Translation
Sorting and degradation
Replication and repair
Regulatory pathways: environmental information processing
Membrane transport
Signal transduction
Ligand-receptor interaction
Regulatory pathways: cellular processes
Cell motility
Cell growth and death
Cell communication
Development
Behavior
Regulatory pathways: human diseases
Neurodegenerative disorders

Source: From ▸http://www.genome.ad.jp/kegg/.

described in bacteria are not always applicable to eukaryotic organisms. Thus KEGG pathways must be used with caution.

Gene expression data from a microarray experiment can be annotated with KEGG pathway numbers and then visualized with expression values mapped onto KEGG pathways. This is offered by the DRAGON database (▸http://pevsnerlab.kennedykrieger.org).

PERSPECTIVE

In the past decade, our understanding of the properties of proteins has advanced dramatically, from the level of biochemical function to the role of proteins in cellular processes. Many web-based tools are available to evaluate the biochemical features of individual proteins. Additionally, high-throughput approaches such as two-dimensional gel electrophoresis and the yeast two-hybrid system have been used in an effort to define the function of all proteins. Large numbers of proteins still have no known function because they lack detectable homology to other characterized proteins. We will continue to obtain a more comprehensive description of protein function as distinct high-throughput strategies are applied to model organisms, such as large-scale analyses of protein localization and protein interactions.

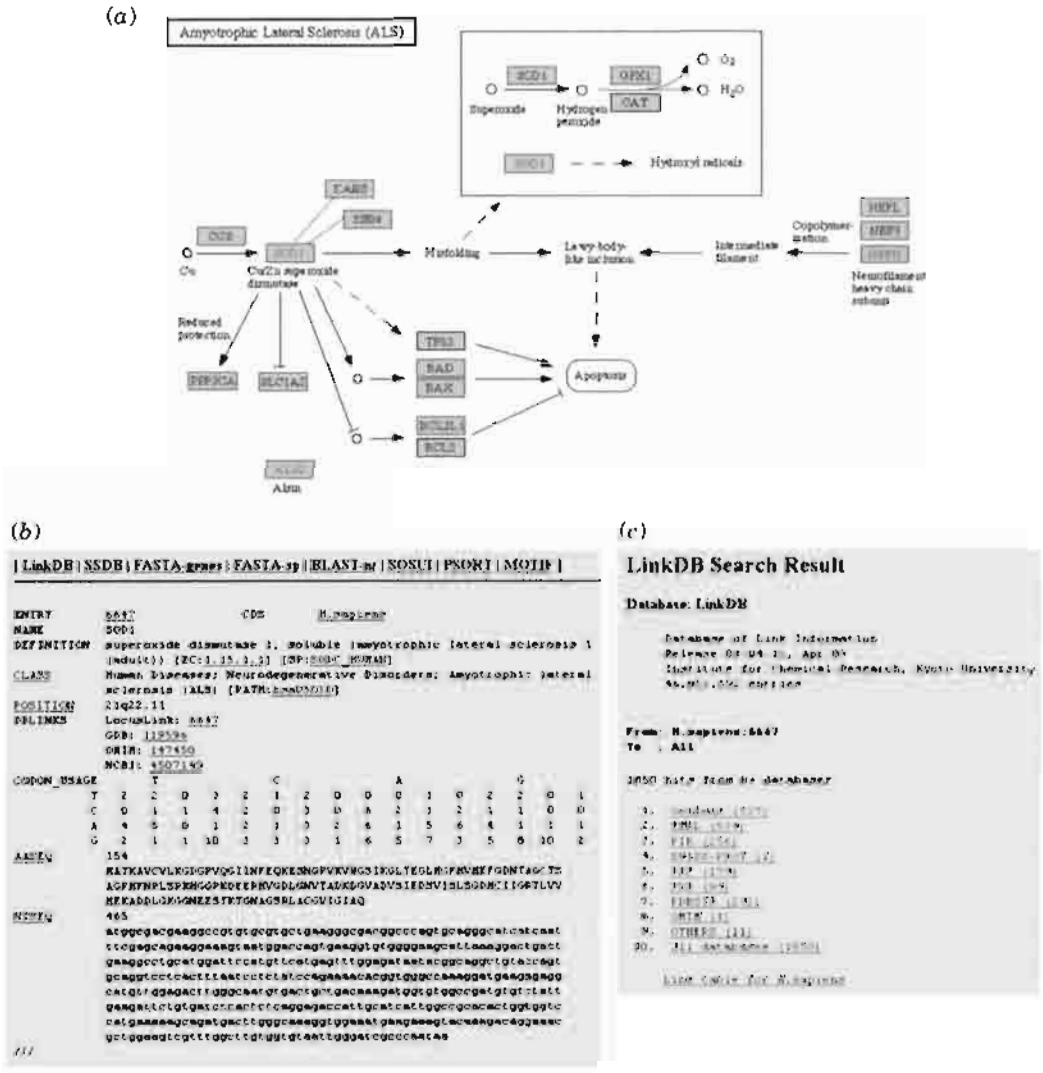


FIGURE 8.30. *The KEGG database includes pathway maps and data on over 350,000 genes and proteins from about 100 completed genomes. One of the pathway maps is for diseases, and the map for amyotrophic lateral sclerosis (ALS; Lou Gehrig's disease) is shown in* (a). *The superoxide dismutase gene* SOD1 *is linked to many familial cases of ALS;* (b) *clicking on its icon leads to the* SOD1 *entry in KEGG as well as* (c) *links to protein prediction tools described in this chapter.*

PITFALLS

Many of the experimental and computational strategies used to study protein function have limitations. Two-dimensional protein gels are most useful for studying relatively abundant proteins, but thousands of proteins expressed at low levels are harder to characterize. The yeast two-hybrid approach may have many false-negative results. It will not identify protein–protein interactions that fail to occur in the yeast nucleus or under the constraints of a binding assay employing recombinant fusion proteins.

Once high-throughput approaches to protein function are used to classify protein function, these assignments should be regarded only as tentative hypotheses. Conventional cell and molecular biology experiments are required to verify and explore protein function experimentally.

WEB RESOURCES

TABLE 8-11 Tools to Analyze Protein Motifs

Program	Comment	URL
InterProScan	At EBI	▸http://www.ebi.ac.uk/interpro/scan.html
ppsearch	At EBI	▸http://www2.ebi.ac.uk/ppsearch/
PRATT	At EBI	▸http://www2.ebi.ac.uk/pratt/
ProfileScan Server	At ISREC	▸http://hits.isb-sib.ch/cgi-bin/hits.motifscan
PROSCAN (PROSITE SCAN)	At PBIL (Pôle Bio-Informatique Lyonnais) (▸http://pbil.univ-lyon1.fr/)	▸http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_prosite.html
ScanProsite tool	At ExPASy	▸http://ca.expasy.org/tools/scnpsite.html
SMART	At EMBL	▸http://smart.embl-heidelberg.de/
TEIRESIAS	At IBM	▸http://cbcsrv.watson.ibm.com/Tspd.html

Source: From ExPASy: ▸http://www.expasy.org/tools/.

TABLE 8-12 Tools to Analyze Primary and/or Secondary Structure Features of Proteins

Program	Source/Comment	URL
COILS	Prediction of coiled-coil regions in proteins	▸http://www.ch.embnet.org/software/COILS_form.html
Compute pI/Mw	From ExPASy	▸http://www.expasy.org/tools/pi_tool.html
drawhca	Hydrophobic cluster analysis plot	▸http://psb00.snv.jussieu.fr/hca/hca-seq.html
Helical Wheel	Draws an helical wheel, i.e., an axial projection of a regular alpha helix, for a given sequence	▸http://www.site.uottawa.ca/~turcotte/resources/HelixWheel/
M.M., pI, composition, titrage	From the Atelier Bio Informatique de Marseille	▸http://www.up.univ-mrs.fr/~wabim/d_abim/compo-p.html
Paircoil	Prediction of coiled-coil regions in proteins	▸http://nightingale.lcs.mit.edu/cgi-bin/score
PeptideMass	From ExPASy	▸http://kr.expasy.org/tools/peptide-mass.html
REP	Searches a protein sequence for repeats	▸http://www.embl-heidelberg.de/~andrade/papers/rep/search.html
SAPS	Statistical analysis of protein sequences	▸http://www.isrec.isb-sib.ch/software/SAPS_form.html

Source: From ExPASy: ▸http://www.expasy.org/tools/.

TABLE 8-13 Web Resources for Characterization of Glycosylation Sites on Proteins

Program	Comment/Source	URL
DictyOGlyc 1.1 Prediction Server	Neural network predictions for GlcNAc O-glycosylation sites in *Dictyostelium discoideum* proteins	▸http://www.cbs.dtu.dk/services/DictyOGlyc/
NetOGlyc	Prediction of type O-glycosylation sites in mammalian proteins	▸http://www.cbs.dtu.dk/services/NetOGlyc/
YinOYang 1.2	Produces neural network predictions for O-ß-GlcNAc attachment sites in eukaryotic protein sequences	▸http://www.cbs.dtu.dk/services/YinOYang/

TABLE 8-14 Tools to Analyze Postranslational Modifications

Program	Comment	URL
big-PI Predictor	GPI modification site prediction	▸http://mendel.imp.univie.ac.at/gpi/
DGPI	Detection/prediction of GPI cleavage site (GPI-anchor) in a protein	▸http://129.194.186.123/GPI-anchor/index.en.html
NetPhos 2.0 Prediction Server	Produces neural network predictions for serine, threonine, and tyrosine phosphorylation sites in eukaryotic proteins	▸http://www.cbs.dtu.dk/services/NetPhos/
Sulfinator	Prediction of tyrosine sulfation sites	▸http://www.expasy.org/tools/sulfinator/

Source: From ExPASy: ▸http://www.expasy.org/tools/.

TABLE 8-15 Examples of Proteins with Unusually High Occurrences of Specific Amino Acids

Amino Acid(s)	Proteins
C	Disulfide-rich proteins; metallothioneins; zinc finger proteins
D, E	Acidic proteins
G	Collagens
H	Hisactophilin; histidine-rich glycoprotein
W, L, P, Y, L, V, M, A	Transmembrane domains
K, R	Nuclear proteins (nuclear localization signals)
N	*Dictyostelium* proteins
P	Collagens; filaments; SH3/WW/EVHI binding sites
Q	Proteins encoded by genes mutated in triplet repeat disorders (Chapter 17)
S, R	Some RNA-binding motifs
S, T	Mucins; oligosaccharide attachment sites
abc**d**efg	Heptad coiled coils (**a** and **d** are hydrophobic residues), e.g., myosins

Source: Modified from Ponting (2001). The hydrophobic residues characteristic of transmembrane helices are from Tanford (1980). Used with permission.

TABLE 8-16 Web-Based Programs for Prediction of Protein Localization

Program	Comment	URL
ChloroP	Predicts presence of chloroplast transit peptides (cTP) in protein sequences	▸http://www.cbs.dtu.dk/services/ChloroP/
MITOPROT	Calculates the N-terminal protein region that can support a mitochondrial targeting sequence and the cleavage site	▸http://www.mips.biochem.mpg.de/cgi-bin/proj/medgen/mitofilter
PSORT	Prediction of protein-sorting signals and localization sites	▸http://psort.nibb.ac.jp/
SignalP	Predicts presence and location of signal peptide cleavage sites in prokaryotes and eukaryotes	▸http://www.cbs.dtu.dk/services/SignalP/
TargetP	Predicts subcellular location of eukaryotic protein sequences	▸http://www.cbs.dtu.dk/services/TargetP/

TABLE 8-17 Web Servers for Prediction of Transmembrane Domains in Protein Sequences

Program	Comment/Source	URL
DAS server		▸ http://www.sbc.su.se/~miklos/DAS/
HMMTOP	Prediction of transmembrane helices and topology of proteins	▸ http://www.enzim.hu/hmmtop/
PredictProtein server	Prediction of transmembrane helix location and topology	▸ http://dodo.cpmc.columbia.edu/predictprotein/
SOSUI	Classification and secondary structure prediction of membrane proteins	▸ http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html
TMpred		▸ http://www.ch.embnet.org/software/TMPRED_form.html
TMHMM (v. 2.0)	Center for Biological Sequence Analysis, Technical University of Denmark	▸ http://www.cbs.dtu.dk/services/TMHMM-2.0/
TopPred2	Topology prediction of membrane proteins	▸ http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html

Source: ExPASy web server.

TABLE 8-18 Web-Based Databases of Protein–Protein and Protein–Ligand Interactions

Database	Comment	URL
BIND (The Biomolecular Interaction Network Database)	Database designed to store full descriptions of interactions, molecular complexes, and pathways	▸ http://www.bind.ca/
Cellzome	–	▸ http://yeast.cellzome.com
DIP (The Database of Interacting Proteins)	–	▸ http://dip.doe-mbi.ucla.edu/main.html
DLRP (Database of Ligand-Receptor Partners)	Database of protein ligand and protein receptor pairs that are known to interact with each other	▸ http://dip.doe-mbi.ucla.edu/dlrp/main.html
FlyBase	See Jacq (2001)	▸ http://fly.ebi.ac.uk:7081
FlyNets	See Jacq (2001)	▸ http:/gifts.univ-mrs.fr/FlyNets
KEGG (Kyoto Encyclopedia of Genes and Genomes)	–	▸ http://www.genome.ad.jp/kegg
GeNet (Gene Networks database)	See Jacq (2001)	▸ http://www.csa.ru:85/Inst/gorb_dep/inbios/genet/genet.htm
STKE (Signal Transduction Knowledge Environment)	See Jacq (2001)	▸ http://www.stke.org
Transfac (Transcription Factor database)	–	▸ http://transfac.gbf-braunschweig.de/TRANSFAC/index.html
YPD, PombePD, WormPD	Proteome databases	▸ http://www.proteome.com/databases/index.html

Source: From ▸ http://dip.doe-mbi.ucla.edu/databases.html. Some of these databases are listed in Schächter (2002).

DISCUSSION QUESTIONS

[8-1] InterPro is an important resource that coordinates information about protein signatures from a variety of databases. When these databases all describe a particular protein family or a particular signature, what different kinds of information can you get? Is the information in InterPro redundant?

[8-2] What are some of the major assumptions of high-throughput screening technologies such as the yeast two-hybrid assay? In general terms, how many binding partners does an average protein have and under what physiological conditions?

PROBLEMS

[8-1] Salmon has a pinkish color, and some lobsters are blue (but turn red when boiled) because a chromophore called astaxanthin binds to a carrier protein called crustacyanin. Examine the protein sequence of crustacyanin from the European lobster *Homarus gammarus*. What are some of its physical properties (e.g., molecular weight, isoelectric point)? Does it have any known domains or motifs that might explain how or why it binds to the chromophore? Use the tools at the ExPASy site. (For more information about this protein, read the article at ExPASy: ▸http://www.expasy.org/spotlight/articles/sptlt026.html.)

[8-2] Evaluate human syntaxin at the ExPASy site. Does it have coiled-coil regions? How many predicted transmembrane domains does it have? What is its function? Use the ExPASy sequence retrieval system first.

[8-3] Olfactory receptors are related to the rhodopsin-like G-protein coupled receptor (GPCR) superfamily. Use the proteome tools at EBI (▸http://www.ebi.ac.uk/proteome/) to decide about what percent of the mouse proteome is comprised of these receptors. About what percent of the human proteome is comprised of these receptors?

[8-4] Explore the proteome tools at ▸http://www.ebi.ac.uk/proteome/. Are any of the 15 most common protein domains in *E. coli* K12 also present in human?

SELF-TEST QUIZ

[8-1] Can a domain be at the amino terminus of one protein and the carboxy terminus of another protein?
(a) Yes
(b) No

[8-2] In general, if you compare the size of a pattern (also called a motif or fingerprint) and a domain:
(a) They are about the same size.
(b) The pattern is larger.
(c) The pattern is smaller.
(d) The comparison always depends on the particular proteins in question.

[8-3] The amino acid sequence [ST]-X-[RK] is the consensus for phosphorylation of a substrate by protein kinase C. This sequence is an example of:
(a) A motif that is characteristic of proteins that are homologous to each other
(b) A motif that is characteristic of proteins that are not necessarily homologous to each other
(c) A domain that is characteristic of proteins that are homologous to each other
(d) A domain that is characteristic of proteins that are not necessarily homologous to each other

[8-4] If you analyze a single, previously uncharacterized protein using programs that predict glycosylation, sulfation, phosphorylation, or other posttranslational modifications:
(a) The predictions of the programs are not likely to be accurate.
(b) The accuracy of the predictions is unknown and difficult to assess.
(c) The predictions of the programs are likely to be accurate concerning the possible presence of particular modifications, but their biological revelance is unknown until you assess the protein's properties experimentally.
(d) The predictions of the programs are likely to be accurate concerning the possible presence of particular modifications, but it is not feasible to assess the protein's properties experimentally.

[8-5] An underlying assumption of the Gene Ontology Consortium is that the description of a gene or gene product according to three categories (molecular function, biological process, and cellular component):
(a) Is likely to be identical across many species, from plants to worms to human
(b) Is likely to vary greatly across many species, from plants to worms to human
(c) May or may not be identical across many species and thus must be assessed for each gene or gene product individually
(d) May or may not be identical across many species and thus must be assessed for each gene or gene product individually by an expert curator

[8-6] Protein localization is described primarily in which Gene Ontology category?
(a) Molecular function
(b) Cellular component
(c) Cellular localization
(d) Biological process

[8-7] Which of the following is a means of assessing protein function?
(a) Finding structural homologs
(b) Studying bait–prey interactions
(c) Determining the isoelectric point
(d) All of the above

[8-8] A major advantage of two-dimensional protein gels as a high-throughput technology for protein analysis is that:
(a) Sample preparation and the process or running two-dimensional gels is straightforward and can be automated.

(b) The result of two-dimensional gels includes data on both the size and the charge of thousands of proteins.
(c) The technique is well suited to the detection of low-abundance proteins.
(d) The technique is well suited to the detection of hydrophobic proteins.

[8-9] High-throughput screens such as the yeast two-hybrid system and affinity purification experiments can have false-positive results because:
(a) Some proteins are inherently sticky.
(b) Some bait proteins that are introduced into cells become mislocalized.
(c) Some protein complexes form only very transiently.
(d) Affinity tags or epitope tags can interfere with protein-protein interactions.

[8-10] Which of the following best describes a major problem in evaluating large-scale cellular pathway diagrams?
(a) The direction of the biochemical pathways is not usually known.
(b) The pathway maps do not employ Gene Ontology nomenclature.
(c) The pathway maps often depend on the correct identification of orthologs, but this can be problematic.
(d) The pathway maps tend to be derived from prokaryotes, but only limited information is available on eukaryotes.

Suggested Reading

There are many excellent articles describing protein families. Of particular interest are papers by Steven Henikoff, Peer Bork, and colleagues (1997) and a review in *Scientific American* by Russell Doolittle and Peer Bork (1993).

Bernard Jacq (2001) has written a superb review on protein function. This article discusses the complexity of protein function and the use of new bioinformatic tools to dissect function. Jacq proposes to consider function from six structural levels, from the structure of a protein to its role in a population of organisms.

References

Apweiler, R., et al. InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* **16,** 1145–1150 (2000).

Ashburner, M., et al. Creating the gene ontology resource: Design and implementation. *Genome Res.* **11,** 1425–1433 (2001).

Ashburner, M., et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).

Austen, B. M., and Westwood, O. M. *Protein Targeting and Secretion.* IRL Press, Oxford, 1991.

Bork, P., and Gibson, T. J. Applying motif and profile searches. *Methods Enzymol.* **266,** 162–184 (1996).

Bork, P., and Koonin, E. V. Protein sequence motifs. *Curr. Opin. Struct. Biol.* **6,** 366–376 (1996).

Cooper, T. G. *The Tools of Biochemistry.* Wiley, New York, 1977.

Copley, R. R., Schultz, J., Ponting, C. P., and Bork, P. Protein families in multicellular organisms. *Curr. Opin. Struct. Biol.* **9,** 408–415 (1999).

Doolittle, R. F. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* **64,** 287–314 (1995).

Doolittle, R. F., and Bork, P. Evolutionarily mobile modules in proteins. *Sci. Am.* **269,** 50–56 (1993).

Dunn, M. J. (ed.) From Genome to Proteome: Advances in the Practice and Application of Proteomics. Wiley-VCH, New York 2000.

Enright, A. J., Iliopoulos, I., Kyrpides, N. C., and Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **402,** 86–90 (1999).

Farmer, T. B., and Caprioli, R. M. Determination of protein-protein interactions by matrix-assisted laser desorption/ionization mass spectrometry. *J. Mass Spectrom.* **33,** 697–704 (1998).

Fields, S., and Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340,** 245–246 (1989).

Fountoulakis, M., Schuller, E., Hardmeier, R., Berndt, P., and Lubec, G. Rat brain proteins: Two-dimensional protein database and variations in the expression level. *Electrophoresis* **20,** 3572–3579 (1999).

Frankel, A. D., and Young, J. A. HIV-1: Fifteen proteins and an RNA. *Annu. Rev. Biochem.* **67,** 1–25 (1998).

Gavin, A. C., et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147 (2002).

Geppert, M., Goda, Y., Stevens, C. F., and Sudhof, T. C. The small GTP-binding protein Rab3A regulates a late step in synaptic vesicle fusion. *Nature* **387,** 810–814 (1997).

Grünenfelder, B., et al. Proteomic analysis of the bacterial cell cycle. *Proc. Natl. Acad. Sci. USA* **98,** 4681–4686 (2001).

Henikoff, S., et al. Gene families: The taxonomy of protein paralogs and chimeras. *Science* **278,** 609–614 (1997).

Ho, Y., et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415,** 180–183 (2002).

Hoogland, C., et al. The SWISS-2DPAGE database: What has changed during the last year. *Nucleic Acids Res.* **27,** 289–291 (1999).

Ito, T., et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. USA* **98,** 4569–4574 (2001).

Jacq, B. Protein function from the perspective of molecular interactions and genetic networks. *Brief. Bioinform.* **2,** 38–50 (2001).

Jones, D.T. Protein structure prediction in genomics. *Brief. Bioinform.* **2,** 111–125 (2001).

Kallioniemi, O. P., Wagner, U., Kononen, J., Sauter, G. Tissue microarray technology for high-throughput molecular profiling of cancer. *Hum. Mol. Genet.* **10,** 657–662 (2001).

Kanehisa, M., and Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

Karp, P. D. Pathway databases: A case study in computational symbolic theories. *Science* **293,** 2040–2044 (2001).

Karp, P. D., et al. The EcoCyc Database. *Nucleic Acids Res.* **30,** 56–58 (2002a).

Karp, P. D., Riley, M., Paley, S. M., and Pellegrini-Toole, A. The MetaCyc Database. *Nucleic Acids Res.* **30,** 59–61 (2002b).

Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M. J., Sauter, G., Kallioniemi, O. P. Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nat. Med.* **4,** 844–847 (1998).

Kumar, A., and Snyder, M. Protein complexes take the bait. *Nature* **415,** 123–124 (2002).

Kumar, A., et al. Subcellular localization of the yeast proteome. *Genes Dev.* **16,** 707–719 (2002).

Langen, H., et al. Two-dimensional map of human brain proteins. *Electrophoresis* **20,** 907–916 (1999).

Li, X. J., et al. A huntingtin-associated protein enriched in brain with implications for pathology. *Nature* **378,** 398–402 (1995).

Lupas, A. Predicting coiled-coil regions in proteins. *Curr. Opin. Struct. Biol.* **7,** 388–393 (1997).

Lupas, A., Van Dyke, M., and Stock, J. Predicting coiled coils from protein sequences. *Science* **252,** 1162–1164 (1991).

MacBeath, G. Protein microarrays and proteomics. *Nat. Genet.* Supplement **32,** 526–532 (2002).

Mann, G. *The Chemistry of the Proteids.* The Macmillan Company, New York, 1906.

Marcotte, E. M., et al. Detecting protein function and protein–protein interactions from genome sequences. *Science* **285,** 751–753 (1999a).

Marcotte, E. M., Pellegrini, M., Thompson, M. J., Yeates, T. O., and Eisenberg, D. A combined algorithm for genome-wide prediction of protein function. *Nature* **402,** 83–86 (1999b).

Molloy, M. P., and Witzmann, F. A. Proteomics: Technologies and applications. *Briefings in Functional Genomics and Proteomics* **1,** 23–39 (2002).

Mulder, N. J., et al. InterPro: An integrated documentation resource for protein families, domains, and functional sites. *Brief. Bioinform.* **3,** 225–235 (2002).

Mulder, N. J., et al. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31,** 315–318 (2003).

O'Farrell, P. H. High resolution two-dimensional electrophoresis of proteins. *J Biol. Chem.* **250,** 4007–4021 (1975).

Østergaard, M., et al. Proteome profiling of bladder squamous cell carcinomas: Identification of markers that define their degree of differentiation. *Cancer Res.* **57,** 4111–4117 (1997).

Paley, S. M., and Karp, P. D. Evaluation of computational metabolic-pathway predictions for *Helicobacter pylori*. *Bioinformatics* **18,** 715–724 (2002).

Pevsner, J., Hou, V., Snowman, A. M., and Snyder, S. H. Odorant-binding protein. Characterization of ligand binding. *J. Biol. Chem.* **265,** 6118–6125 (1990).

Ponting, C. P. Issues in predicting protein function from sequence. *Brief. Bioinform.* **2,** 19–29 (2001).

Rain, J. C., et al. The protein–protein interaction map of *Helicobacter pylori*. *Nature* **409,** 211–215 (2001).

Ratnam, M., Nguyen, D. L., Rivier, J., Sargent, P. B., and Lindstrom, J. Transmembrane topography of nicotinic acetylcholine receptor: Immunochemical tests contradict theoretical predictions based on hydrophobicity profiles. *Biochemistry* **25,** 2633–2643 (1986).

Sanchez, J. C., et al. The mouse SWISS-2D PAGE database: A tool for proteomics study of diabetes and obesity. *Proteomics* **1,** 136–163 (2001).

Schächter, V. Bioinformatics of Large-Scale Protein Interaction Networks. *Computational Proteomics. A Supplement to BioTechniques,* 16–27 (2002).

Sigrist, C. J., et al. PROSITE: A documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **3,** 265–274 (2002).

Sonnhammer, E. L., and Kahn, D. Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.* **3,** 482–492 (1994).

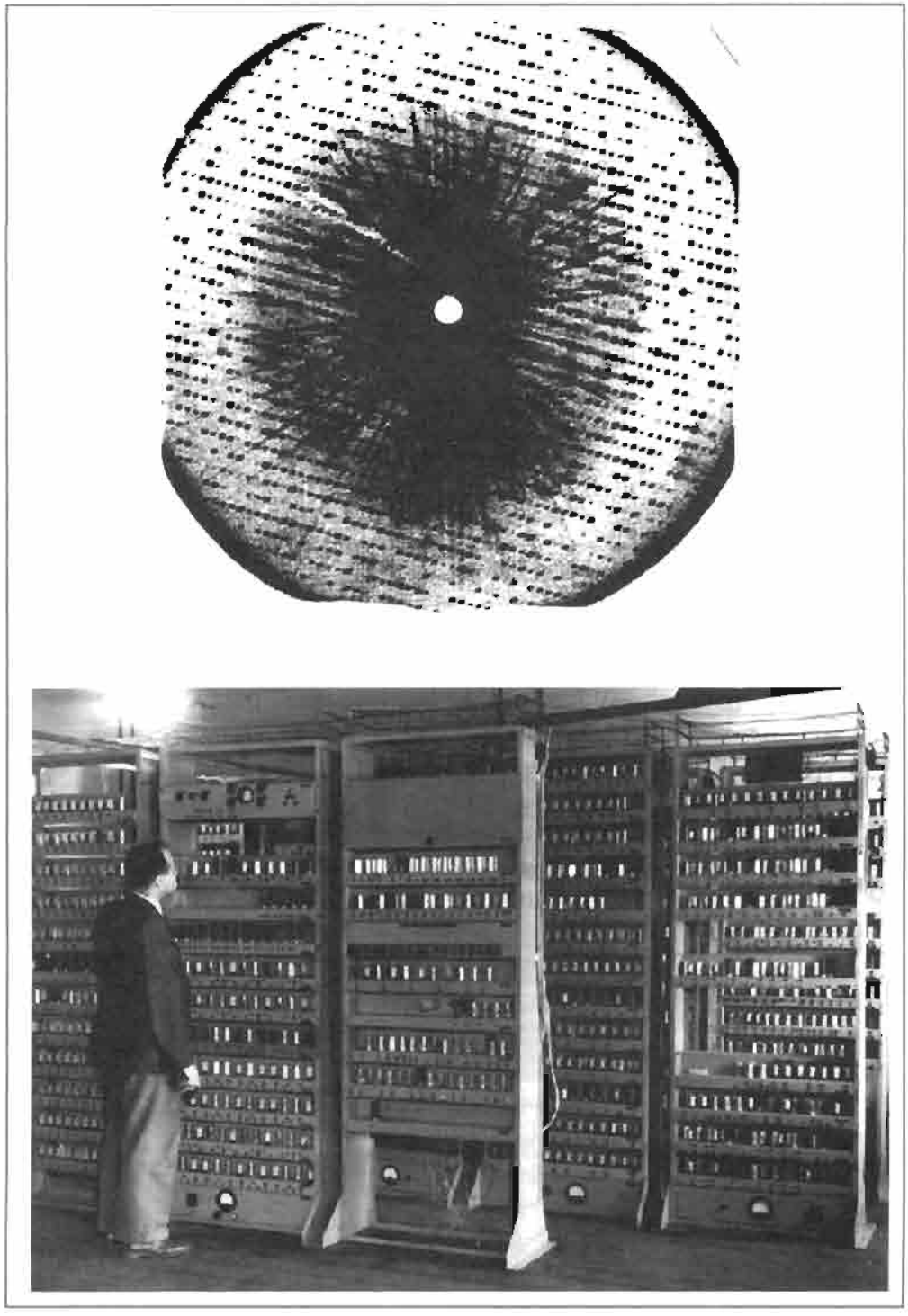
Stroud, R. M., and Walter, P. Signal sequence recognition and protein targeting. *Curr. Opin. Struct. Biol.* **9,** 754–759 (1999).

Takai, Y., Sasaki, T., and Matozaki, T. Small GTP-binding proteins. *Physiol. Rev.* **81,** 153–208 (2001).

Tanford, C. *The Hydrophobic Effect: Formation of Micelles and Biological Membranes.* John Wiley & Sons, New York, 1980.

Uetz, P., et al. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403,** 623–627 (2000).

This Page Intentionally Left Blank



Beginning in the 1940s, Max Perutz and John Kendrew realized the goal of determining the structure of globular proteins by solving the structure of hemoglobulin and myoglobin. In recognition of this work, they shared the Nobel Prize in Chemistry in 1962. (top) *X-ray precession photograph of a myoglobin crystal (from* ▸ *http://www.nobel.se/chemistry/laureates/1962/kendrew-lecture.pdf). Kendrew studied myoglobin from the sperm whale* (Physeter catodon), *and incorporated a heavy metal by the method of isomorphous replacement. He could then bombard the crystals with X-rays in order to obtain an X-ray diffraction pattern (such as that shown here) with which to deduce the electron density throughout the crystal. This required the analysis of 25,000 reflections.* (bottom) *Perutz and Kendrew used the EDSAC I computer (introduced in 1949 and shown here from* ▸ *http://www.cl.cam.ac.uk/Relics/jpegs/edsac99.36.jpg). This computer was essential to interpret the diffraction patterns. For a simulator that shows the capacity of the EDSAC machine, see* ▸ *http://www.dcs.warwick.ac.uk/~edsac/.*

9

Protein Structure

Overview of Protein Structure and Structural Genomics

The three-dimensional structure of a protein determines its capacity to function. High-resolution protein structures have now been obtained for about 20,000 proteins. Traditionally, researchers obtained the structure of individual proteins by starting with information about the known function of the protein (Fig. 9.1). The new approach of structural genomics is based upon a reverse strategy: Genome sequence projects generate predictions of protein-coding sequences (Chapters 12–17). One fundamentally important aspect of each predicted protein is its structure. Predicted proteins may be expressed and their structures are solved (Fig. 9.1). The recent identification of hundreds of thousands of novel predicted proteins has enabled researchers to choose structures to solve based upon a variety of criteria (see below).

The long-term goal of the field of structural genomics is to solve structures that span the full extent of sequence space (Thornton et al., 2000; Burley et al., 1999; Burley, 2000: Koonin et al., 2002). This space may be defined in terms of protein sequence families, which contain members having greater than about 30% amino acid identity. Thus structural genomics aims to solve at least one high-resolution structure for every sequence family. This goal is in some ways as broad as the goal of the Human Genome Project to sequence the entire human genome and will be costlier and take more time to complete.

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

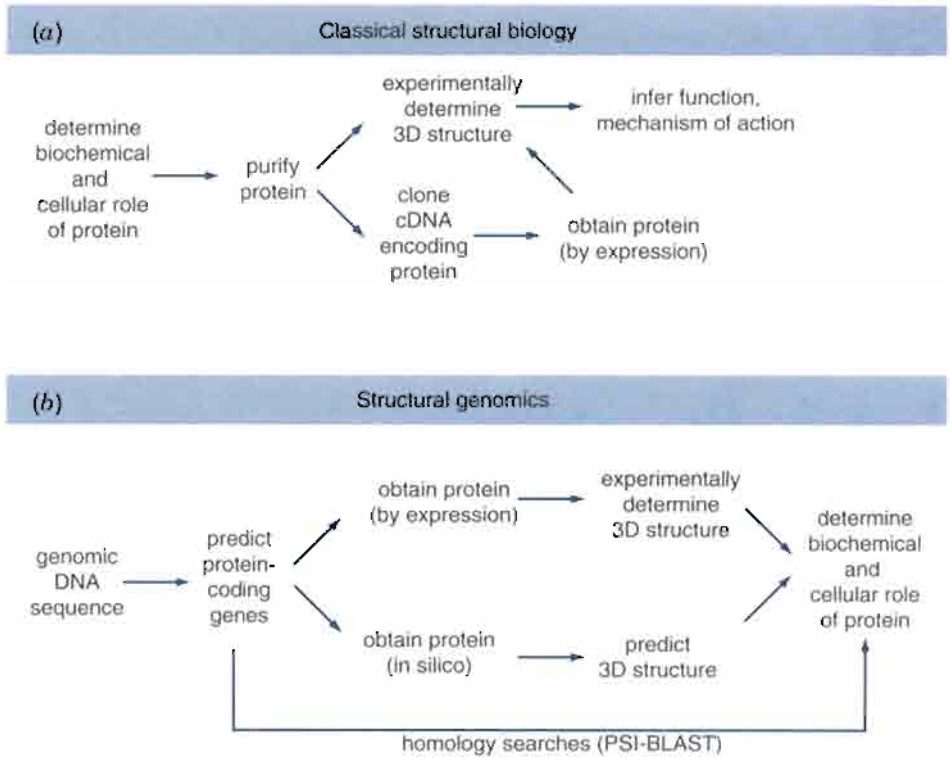


FIGURE 9.1. *(a) In classical structural biology approaches, a protein is purified based upon some known function or activity. After biochemical purification of the protein, if there is sufficient yield, the protein may be crystallized and its structure determined. This in turn allows one to study the biochemical function of the protein and its mechanism of action. Having obtained protein sequence, the corresponding cDNA may be cloned, allowing recombinant protein to be expressed and purified for structure analyses. (b) The classical approach contrasts with that adopted by the field of structural genomics, which proceeds from genomic DNA sequence. Protein-coding genes are predicted, and proteins are either cloned and expressed for biochemical analysis or analyzed by computer-based approaches ("in silico"). The three-dimensional (3D) structure of a protein is then determined experimentally using techniques such as X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, and the structure is predicted computationally. Finally, the biochemical role is determined based upon the nature of the structure. Additional insight into biochemical function is derived from database searches of the protein sequence (e.g., using PSI-BLAST).*

Two proteins that share a very similar structure are usually assumed to also share a similar function. For example, two receptor proteins may share a very similar structure, and even if they differ in their ability to bind ligands or transduce signals, nonetheless they still share the same basic function.

Protein Structure, Homology, and Functional Genomics

As described in Chapter 8, one of the most fundamental questions about a protein is its function. Function is often assigned based upon homology to another protein whose function is perhaps already known or inferred (Holm, 1998; Domingues et al., 2000).

Various types of BLAST searching are employed to identify such relationships of homology (Chapters 4 and 5). However, for many proteins sequence identity is extremely limited. We may take retinol-binding protein and odorant-binding protein as examples: These are both lipocalins of about 20 kDa and are abundant, secreted carrier proteins. They share a GXW motif that is characteristic of lipocalins. However, it is difficult to detect homology based upon analysis of the primary amino acid sequences. By pairwise alignment the two proteins share less than 20% identity. Both structure and function are preserved over evolutionary time more than is sequence identity. Thus the three-dimensional structures of these proteins are extraordinarily similar.

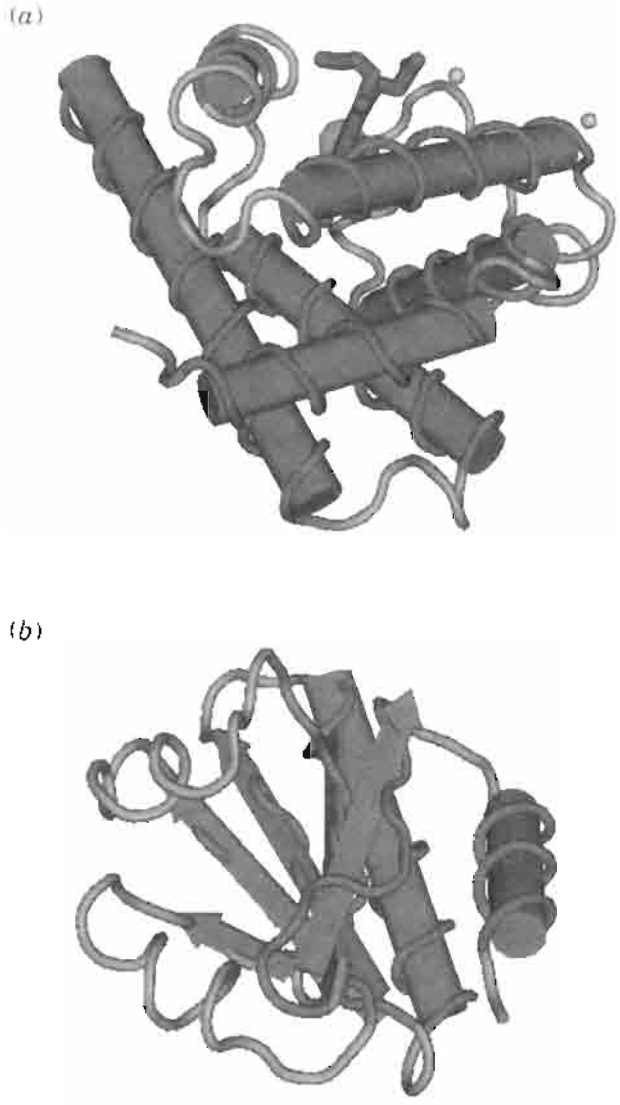


FIGURE 9.2. *Examples of secondary structure. (a) Myoglobin (Protein Data Bank ID 4MBN) is composed of large regions of α helices, shown as strands wrapped around barrel-shaped objects. (b) Thioredoxin from E. coli (PDB 1TXX) is an example of a protein with five β strands, drawn as five large arrows. The most common situation is for proteins to have both α helices and β sheets. Protein taxonomy is described in databases such as CATH (see below).*

Can we generalize about the relationship between amino acid sequence identity and protein structures? It is clear that even a single amino acid substitution can cause a dramatic change in protein structure, as exemplified by disease-causing mutations (discussed at the end of this chapter). Many other substitutions have no observable effects on protein structure, and as in the case of lipocalins, it is common for amino acid sequence to change more rapidly than three-dimensional structure. Wood and Pearson (1999) examined 36 protein families, each having five or more members with known three-dimensional structures. They found a very high, linear correlation between sequence similarity and structural similarity for three-quarters of the protein families. Wood and Pearson concluded that most amino acid sequence changes cause detectable structural changes. Also, the amount of structural change is relatively constant within a protein family.

A goal of structural bioinformatics is to determine the three-dimensional structure of all the major protein families throughout the tree of life (Gerstein and Levitt, 1997; Wolf et al., 1999; Koonin et al., 2002). This will permit a deeper understanding

It is difficult to make a pairwise alignment of rat retinol-binding protein (P04916) and rat odorant-binding protein (NP_620258). If you use BLAST 2 Sequences, no significant match is found, even using a large expect value and a scoring matrix appropriate for distantly related proteins (PAM250). If you do a PSI-BLAST search with rat OBP as a query, you will eventually detect retinol-binding protein after many iterations.

TABLE 9-1 Secondary-Structure Prediction Programs Available on Internet

Program	Comment	URL
APSSP	Based on neural networks	http://imtech.ernet.in/raghava/apssp/
GOR4	From the Pole Bio-Informatique Lyonnais	http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html
Jpred	From the Barton group (Dundee)	http://www.compbio.dundee.ac.uk/~www-jpred/
NNPREDICT	An enhanced neural network approach (from UCSF)	http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html
PHD	Based on neural networks	http://cubic.bioc.columbia.edu/predictprotein
Predator	From the Argos group	http://www-db.embl-heidelberg.de/jss/servlet/de.embl.bk.wwwTools.GroupLeftEMBL/argos/predator/predator_info.html
PredictProtein server	From Columbia University	http://cubic.bioc.columbia.edu/predictprotein/
PSIPRED	From the University College London	http://bioinf.cs.ucl.ac.uk/psipred/
SAM-T99sec	Uses hidden Markov models (Chapter 10)	http://www.cse.ucsc.edu/research/compbio/HMM-apps/T99-query.html
Sosui	From the Mitaku Group (Tokyo)	http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html

Note: Additional sites are listed at ExPASy (http://kr.expasy.org/tools/#secondary), PBIL (http://npsa-pbil.ibcp.fr), and EVA (http://cubic.bioc.columbia.edu/eva/).

of the relatedness of protein domains. This will also enable us to assign function to many proteins.

Biological Questions Addressed by Structural Genomics: Lipocalins

We can use the lipocalin family of transporters to illustrate some of the key questions in structural biology:

- What ligand does each protein transport? For many the answer is unknown. Can structural studies reveal the binding domain to suggest the identity of the ligand? How much structural information is required in order to predict the ligand from sequence information?
- Mutations in some lipocalin genes result in a variety of human diseases. Can we predict the structural and functional consequences of a specific mutation?
- Lipocalins have been divided into subgroups based upon phylogenetic analyses and based upon their localization. To what extent do those groupings reflect structural and functional similarities?
- When a genome is sequenced and a gene encoding a putative novel lipocalin is discovered, can we use information about other lipocalins of known structure in order to predict a new structure?

Principles of Protein Structure: From Primary to Secondary Structure

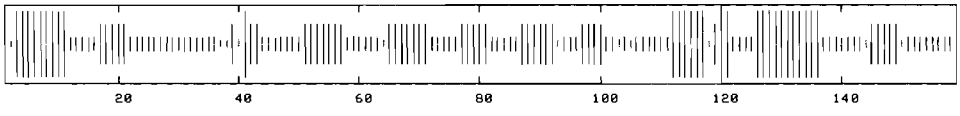
Protein structure is defined at several levels. Primary structure refers to the linear sequence of amino acid residues in a polypeptide chain. The amino acids are joined by peptide bonds on each side of the Cα carbon atom. Secondary structure refers to the arrangements of the primary amino acid sequence into motifs such as α helices, β sheets, and coils (or loops). The tertiary structure is the three-dimensional

(*a*)

```
                  10        20        30        40        50        60        70
                   |         |         |         |         |         |         |
UNK_266410 AQEEEAEQNLSELSGPWRTVYIGSTNPEKIQENGPFRTYFRELVFDDEKGTVDFYFSVKRDGKWKNVHVK
DPM        ccchhhhhhchcctctccceeeeettctcccchtcctccccceehhhehhthccceeeeeeettttcctcceeeh
DSC        ccccccccceeeeccceeeeeeeccccceeeeccccceeeeeeeeccccccceeeeeeeeccccceeeeeeee
GOR4       ccccchhhhhhccccceeeeeeccccccccccccceeeeeeeeccccccceeeeeeeeccccceeeeeeee
HNNC       cchhhhhhhhhhhccccceeeeeeccccccccccccceheeheeeccccccceeeeeeeeccccceeeeee
PHD        ccccccccccccccceeeeeeeeccccccccccccceeeeeeeeeeccccceeeeeeeeccccceeeeee
Predator   cccchhhhhhhccccccceeeeeccccccccccccchhhhhhhccccccceeeeeeeccccccceeeeee
SIMPA96    chhhhhhhhhhhccccceeeeeccccccccccchhhhhhheeccccccceeeeeeeccccccceeeeee
SOPM       hhhhhhhhhhhhhtccceeeeeeccctttcctttcccchhhhheeeettttceeeeeeeectttcceeeeee
Sec.Cons.  cc?hhhhhhhhhccccceeeeeccccccccccccc??e?heeccccccceeeeeeeeccccc?eeeee

                  80        90        100       110       120       130       140
                   |         |         |         |         |         |         |
UNK_266410 ATKQDDGTYVADYEGQNVFKIVSLSRTHLVAHNINVDKHGQTTELTELFVKLNVEDEDLEKFWKLTEDKG
DPM        hhttttcccehctctcceeeeeeeeeeheehhcetetccccchhhheehhhhhhhhhhhhhhhhtcc
DSC        eeccccceeeeeccccceeeeeeeccceeeeeeeeeeeccccccceeeeeeeeccccccchhhhheeccce
GOR4       eeccccceeeeccccccceeeeeccchhhhhccccccccccchhhhhhhhhhhhcchhhhhhhhhhccc
HNNC       eeccccceeeeccccccceeeeeccccceeeeccccccccccchhhhhhheeeccccchhhhhhhhhhccc
PHD        eeeccccceeeeeeeeeeeeeeccceeeeeeeeccccceeeeeeeeccccchhhhhhhhhhhhcc
Predator   eeeccccceeeeccccceeeeeeccccccccccccccchhhhhhhhhhhcccchhhhhhhhhhccc
SIMPA96    ccccccceeeeccccceeeeeccccceeeeccccccccchhhhhhhhhhhcccchhhhhhhhhccc
SOPM       eccttteeeeeettcceeeeeeeccceeeeeeeeeecttcccchhhhhheeeeccchhhhhhhhhccctt
Sec.Cons.  eeccccceeeeeccccceeeeee?cccceeeccc?c?ccccchhhhh?ehhcccchhhhhhhhhccc

                  150
                   |
UNK_266410 IDKKNVVNFLENEDHPHPE
DPM        ctcccceehhhtcctcccc
DSC        ccceeeeeeeeccccccc
GOR4       cccceeeeeeccccccee
HNNC       ccccceeeeeeccccccc
PHD        ccccchhhceccccccce
Predator   ccccccceeeccccccccc
SIMPA96    cchhhhhhhhhccccccc
SOPM       cchhhhhhhhhttccccc
Sec.Cons.  cccc?eeeee?ccccccc

Sequence length :   159
```

(*b*)

```
Sec.Cons. :
   Alpha helix      (Hh) :    29 is  18.24%
   3(10) helix      (Gg) :     0 is   0.00%
   Pi helix         (Ii) :     0 is   0.00%
   Beta bridge      (Bb) :     0 is   0.00%
   Extended strand  (Ee) :    43 is  27.04%
   Beta turn        (Tt) :     0 is   0.00%
   Bend region      (Ss) :     0 is   0.00%
   Random coil      (Cc) :    76 is  47.80%
   Ambigous states  (?)  :    11 is   6.92%
   Other states          :     0 is   0.00%
```



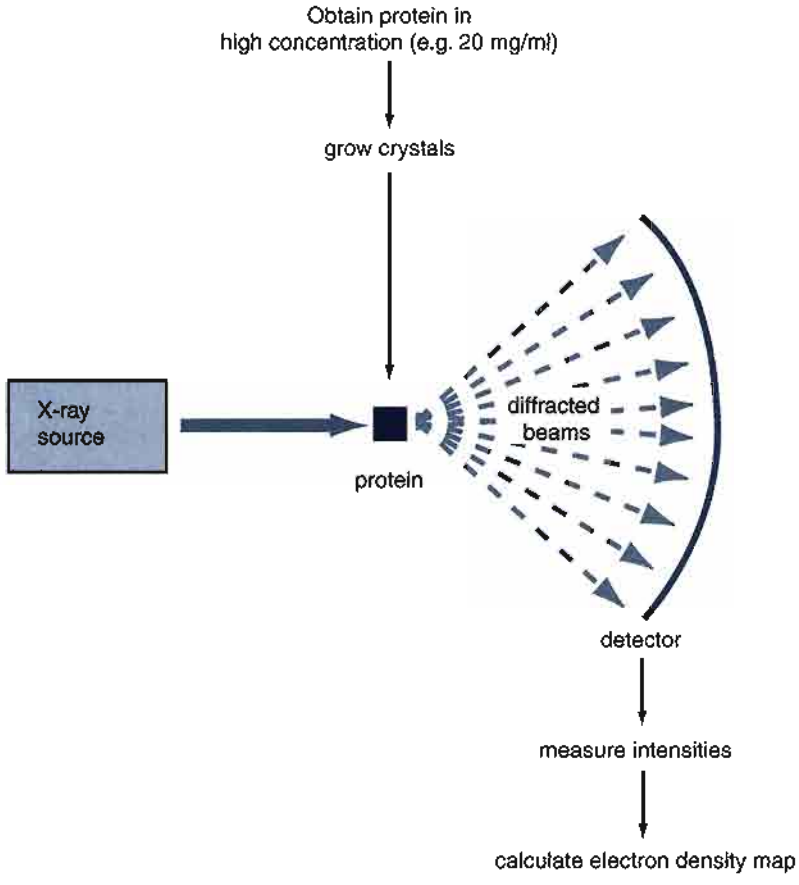

```
PREDATOR parameters :
   Secondary structure data : dssp

SOPM parameters :
   Window width          :  17
   Similarity threshold  :   8
   Number of states      :   4
```

View prediction result file in MPSA : [DPM] [DSC] [GOR4] [HNN] [PHD] [PREDA] [SIMPA96] [SOPM]
View prediction result file in ANTHEPROT : [DPM] [DSC] [GOR4] [HNN] [PHD] [PREDA] [SIMPA96] [SOPM]

FIGURE 9.3. *A variety of web servers offer secondary-structure prediction. The sequence of a lipocalin (bovine odorant-binding protein, accession P07435) was entered in the Network Protein Sequence Analysis site of the Pôle Bio-Informatique Lyonnais. (a) Secondary-structure predictions such as α helices are shown as well as (b) the combined results of nine prediction algorithms. Note that these algorithms offer slightly differing predictions. The letters c, e, h, and t (panel a) are defined in panel b.*

BOX 9-1
X-Ray Crystallography

A protein is obtained in high concentration and crystallized in a solution such as ammonium sulfate. A beam of X rays is aimed at the protein crystals. The protein is in a highly regular array that causes the X rays to diffract (scatter) where they are detected on X-ray film. Spot intensities are measured, and an image is generated by Fourier transformation of the intensities. An electron density map is generated corresponding to the arrangements of the atoms that comprise the protein. Individual atoms are separated by 1–1.5 Å, and resolution of less than 2 Å is generally required for a detailed structure determination.



arrangement formed by packing secondary structure elements into globular domains. Finally, quaternary structure involves this arrangement of several polypeptide chains. Functionally important areas of a protein such as ligand-binding sites or enzymatic active sites are formed at the levels of tertiary and quaternary structure.

In nature, the primary amino acid sequence specifies a three-dimensional structure that forms for each protein. That structure may depend upon some posttranslational modifications, such as the addition of sugars or disulfide bridges. But each cell interprets the information in primary amino acid sequence to form an appropriate structure. The challenge to structural biologists is that we have only

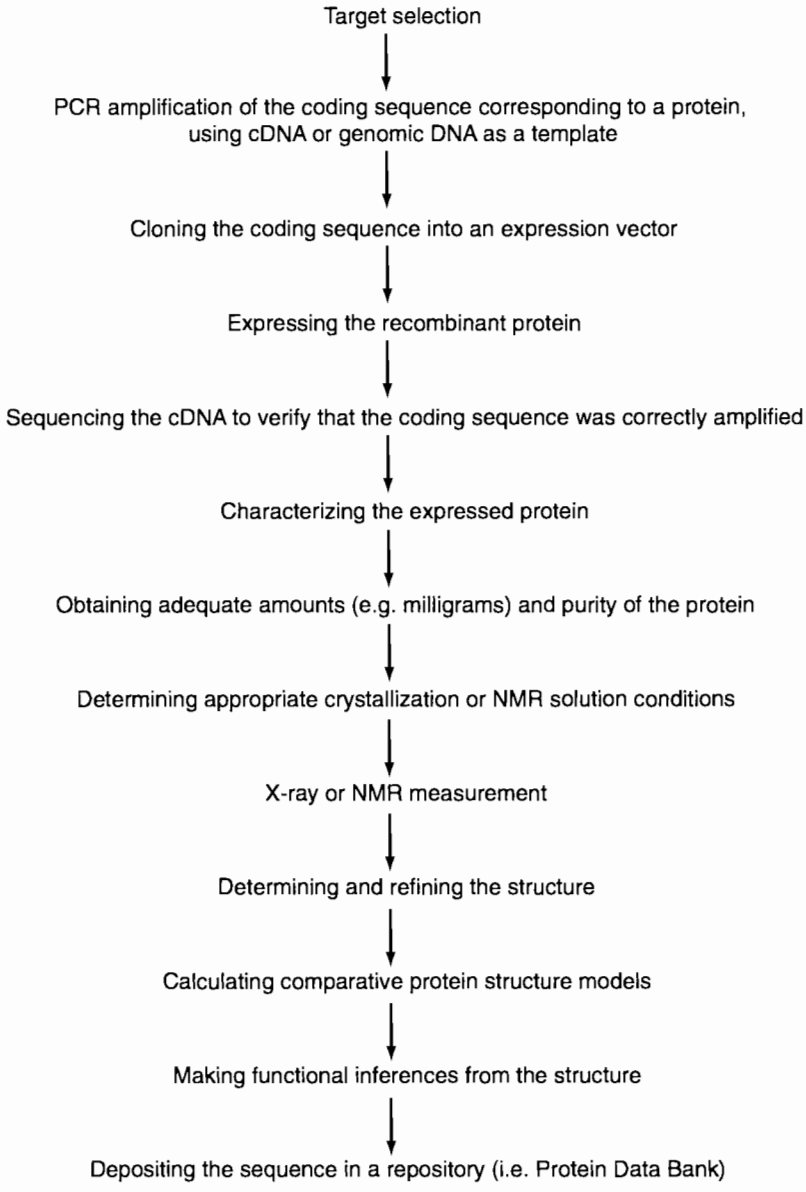


FIGURE 9.4. *General procedure for obtaining a three-dimensional protein structure (modified from Burley, 2000).*

a vague understanding of how to predict a three-dimensional structure based on primary sequence data alone.

Secondary structure is determined by the amino acid side chains. An example of a protein with α helices is myoglobin (Fig. 9.2). The α helices typically are formed from contiguous stretches of 4–40 amino acid residues in length. The β sheets are formed from adjacent β strands composed of 5–10 residues. They are arranged in parallel or antiparallel orientations. An example is shown for thioredoxin from *Escherichia coli* (Fig. 9.2). Most proteins contain combinations of both α helices and β sheets.

Relative to tertiary structure, secondary-structure predictions are easier to make. In the 1970s, Chou and Fasman (1978) developed a method to predict secondary structure based on the frequencies of residues found in α helices, β sheets, and turns. Their algorithm calculates the propensity of each residue to form part of a

FIGURE 9.5. *The PDB is the main repository for three-dimensional structures of proteins and other macromolecules (Berman et al., 2000) (▸ http://www.rcsb.org/pdb). The search fields link allows you to query the PDB in many ways (arrow 1), including with a FASTA search.*

helix, strand, or coil in the context of a sliding window of amino acids. For example, a proline is extremely unlikely to occur in an α helix, and it is often positioned at a turn. The Chou–Fasman algorithm scans through a protein sequence and identifies regions where at least four out of six contiguous residues have a score for α helices above some threshold value. The algorithm extends the search in either direction. Similarly, it searches for bends and turns.

Subsequently other approaches have been developed such as the GOR method of Garnier, Osguthorpe, and Robson (1978) (Garnier et al., 1996). In most cases, these algorithms were used to analyze individual sequences (and they are still useful for this purpose). As multiply aligned sequences have become increasingly available, the accuracy of related secondary-structure prediction programs has increased. The PHD program (Rost and Sander, 1993a,b) is an example of an algorithm that uses multiple sequence alignment for this purpose. The accuracy of the various algorithms has been assessed by evaluating their performance using databases of known structures. Typically, the more recently developed algorithms have about 70–75% accuracy (Rost, 2001; Przybylski and Rost, 2002). This accuracy far exceeds that of the Chou–Fasman algorithm.

TABLE 9-2 PDB Holdings List (June 2003)

Experimental Technique	Molecule Type: Proteins, Peptides, Viruses	Protein and Nucleic Acid Complexes	Nucleic Acids	Carbohydrates	Total
X-ray diffraction and other	16,452	788	673	14	17,927
NMR	2,609	91	520	4	3,224
Total	19,061	879	1,193	18	21,151

Source: From ▸http://www.rcsb.org/pdb/holdings.html.

A variety of web servers allow you to input a primary amino acid sequence and predict the secondary structure (Table 9.1). Some of the programs allow you to enter a single sequence, while others allow you to enter a multiple sequence alignment. As an example, the Pôle Bio-Informatique Lyonnais (PBIL) has a web server that offers secondary-structure predictions for a protein query such as bovine odorant-binding protein (Fig. 9.3*a*). This server also generates predictions using nine different algorithms and calculates a consensus (Fig. 9.3*b*). The various predictions differ somewhat in detail but are generally similar.

The PBIL website is at ▸http://npsa-pbil.ibcp.fr.

Tertiary Protein Structure: Protein-Folding Problem

How does a protein fold into a three-dimensional structure? This problem is solved very rapidly in nature. Often, proteins fold spontaneously almost immediately after protein synthesis. Some proteins only fold after they have traversed the lipid bilayer of the endoplasmic reticulum (ER), where they then fold properly in the lumen of the ER.

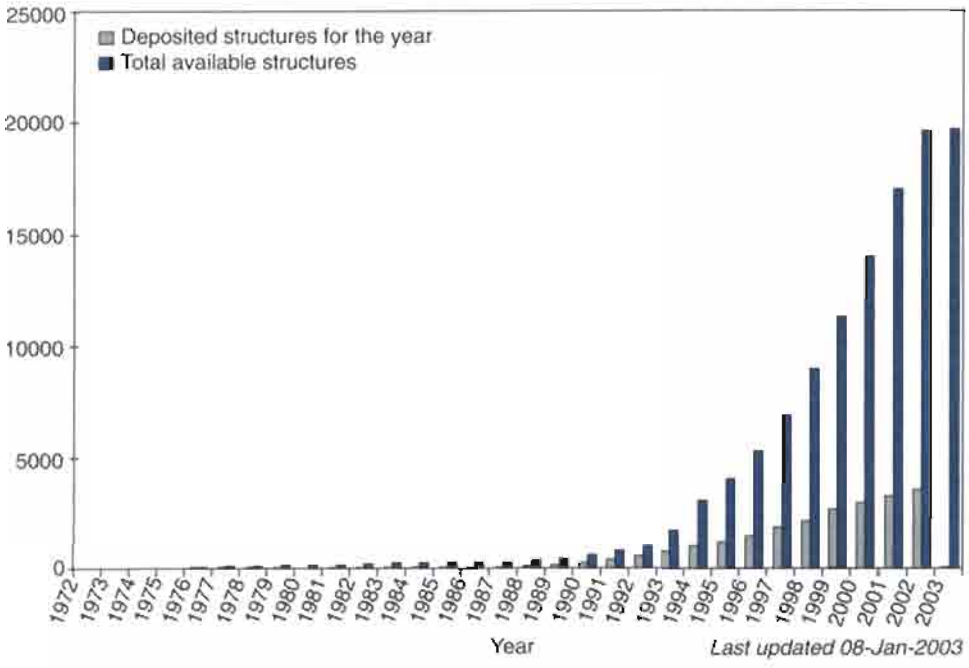


FIGURE 9.6. *The PDB database has grown dramatically in the past decade. The current overview of structures in the database is available at* ▸*http://www.rcsb.org/pdb/holdings.html.*

FIGURE 9.7. *Result of a PDB query for lipocalins.*

In structural biology, there are three main approaches to determining protein structure.

1. Structures can be determined experimentally, most often by X-ray crystallography and nuclear magnetic resonance (NMR).
2. Comparative homology modeling is used to predict the structure of a given protein sequence (the target) by comparison to one or more homologous, known structures (templates) (Marti-Renom et al., 2000; Shortle, 2000). Comparative modeling is most successful when the percent amino acid identity between the target and the template is high.
3. Ab initio methods use physical principles alone (and not template structures) to predict the three-dimensional structure of a target (Osguthorpe, 2000; Simons et al., 2001).

In this chapter we will briefly describe experimental approaches to the determination of protein structure. We will then describe the Protein Data Bank, the principal repository for protein structures. We will show how to use NCBI, ExPASy, and

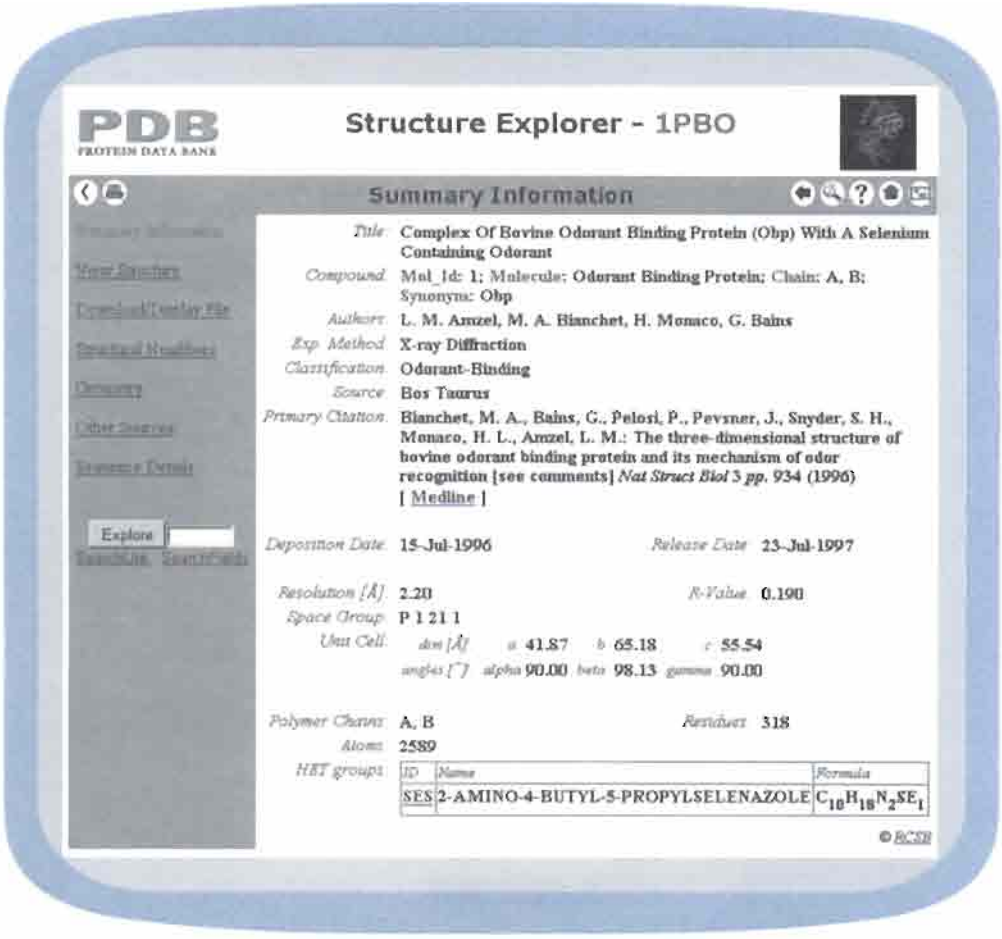


FIGURE 9.8. *Result of a search for the structure of the odorant-binding protein (OBP), a lipocalin. The summary information includes a description of the resolution (2.20 Å), the space group, and the unit cell dimensions. This protein was crystallized in the presence of a derivative of the heavy metal atom selenium. The left sidebar includes links to view the structure of the protein and other information.*

other tools to visualize such structures. A variety of databases such as CATH, SCOP, and Dali classify protein structures. After discussing these databases, we will end the chapter by examining the two main computational methods to protein structure prediction: comparative modeling and ab initio prediction.

Experimental Approaches to Protein Structure

There are two principal approaches to experimental determination of a protein structure: X-ray crystallography and NMR. X-ray crystallography is the most rigorous experimental technique used to determine the structure of a protein (Box 9.1), and about 80% of known structures were determined using this approach. A protein must be obtained in high concentration and seeded in conditions that permit crystallization. The crystal scatters X rays onto a detector, and the structure of the crystal is inferred from the diffraction pattern. The wavelength of X rays (about 0.5–1.5 Å) is useful to measure the distance between atoms, making this technique suitable to trace the amino acid side chains of a protein. The earliest protein structure to be solved was myoglobin (by John Kendrew in 1958) (Fig. 9.2*a*). This revealed a

FIGURE 9.9. *The View Structure page of a PDB entry provides still images of the OBP structure as well as a variety of interactive display options: VRML, Rasmol, MICE, FirstGlance, and Protein Explorer.*

highly irregular shape that is reflective of the complex spatial arrangements that may be adopted by amino acids. In general, proteins tend to be arranged with hydrophobic amino acids in the interior and hydrophilic residues exposed to the surface. This hydrophobic core is produced in spite of the highly polar nature of the peptide backbone of a protein. The most common way that a protein solves this problem is to organize the interior amino residues into secondary structures consisting of α helices and β sheets. While it is possible to predict secondary-structure features, it is far more difficult to predict the way those features are folded into three-dimensional structures.

Nuclear magnetic resonance spectroscopy is an important alternative approach to crystallography. A magnetic field is applied to proteins in solution, and characteristic chemical shifts are observed. From these shifts, the structure is deduced. The largest structures that have been determined by NMR are about 350 amino acids ($\approx$40 kD), considerably smaller than the size of proteins routinely studied by

TABLE 9-3 Interactive Visualization Tools for Protein Structures

Tool	Comment	URL
Chime	Plug-in for a web browser	Instructions at PDB
Cn3D	From NCBI	▸http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
Mage	Reads Kinemages	▸http://www.ncbi.nlm.nih.gov/Structure/CN3D/mage.html
MICE Java applet		Instructions at PDB
RasMol	A stand-alone package	Instructions at PDB
SwissPDB viewer	At ExPASy	▸http://www.expasy.org/spdbv/
VMD	Visual Molecular Dynamics; University of Illinois	▸http://www.ks.uiuc.edu/Research/vmd/
VRML	Uses MolScript	Instructions at PDB

Note: For Protein Data Bank URLs, see the help document at ▸http://www.rcsb.org/pdb/help-graphics.html.

crystallography. However, an advantage of NMR is that it does not require a protein to be crystallized, a notoriously difficult process.

Target Selection and Acquisition of Three-Dimensional Protein Structures

The general procedure for experimentally acquiring protein structural data is outlined in Figure 9.4. This begins with target selection, the process of choosing which structure to solve (Brenner, 2000). Historically, proteins such as hemoglobin and cytochrome *c* were selected that were most amenable to experimental study: They are generally small, soluble, abundant, and known to have interesting biological

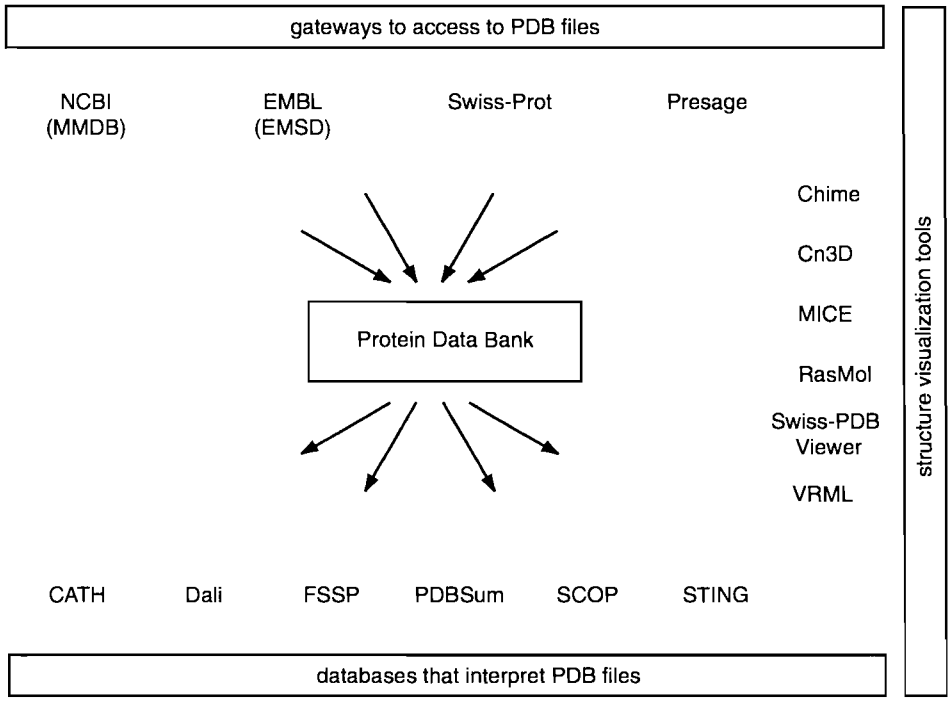


FIGURE 9.10. *The PDB is the central repository for all structural data on proteins and other biological macromolecules. It is housed at ▸http://www.rcsb.org/pdb/. A variety of sites allow access to PDB data, including NCBI and EMBL. Also, many databases analyze PDB structures to generate classification schemes for all protein folds and for other levels of analysis of protein structures. Examples of these databases are SCOP, CATH, Dali, and FSSP (see below). Additional databases are described in Web Resources (Table 9.6).*

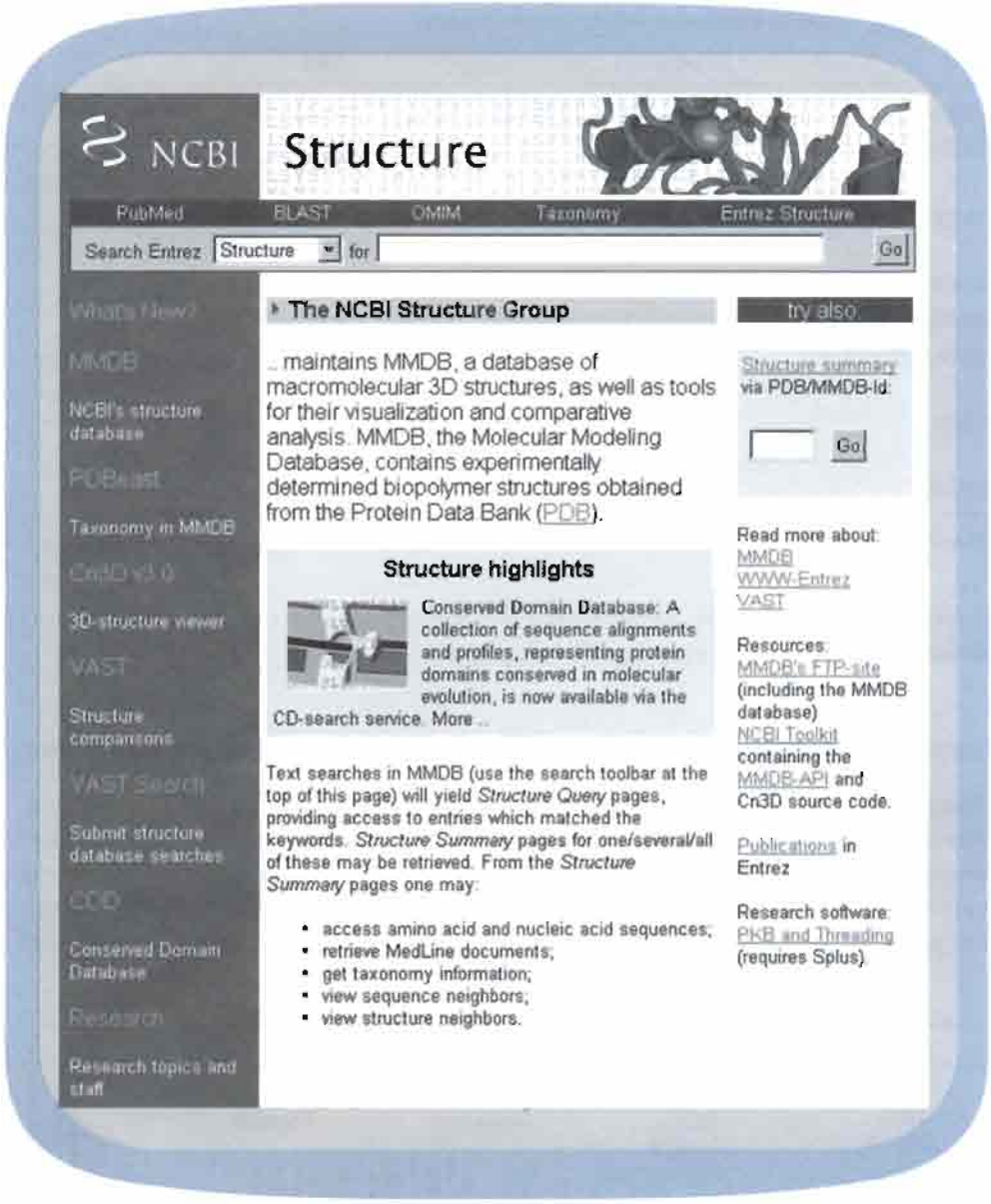
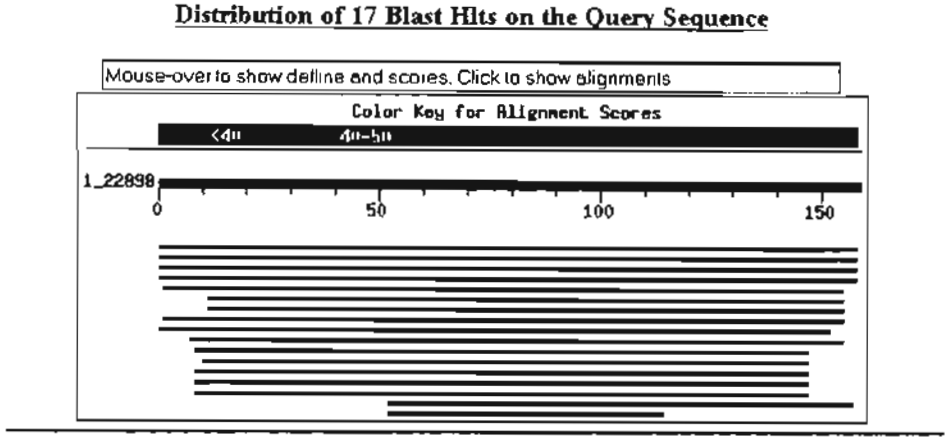


FIGURE 9.11. *The structure site at NCBI offers links to PDB and to tools for structural genomics such as Cn3D (a structure viewer), VAST (a tool to compare structures), and the structure database MMDB (Molecular Modeling DataBase). This site is at* http://www.ncbi.nlm.nih.gov/Structure/.

functions. Today, additional criteria are considered in deciding priorities for which protein structures to solve (McGuffin and Jones, 2002):

- All branches of life (eukaryotes, bacteria, archaea, and viruses) are studied.
- Should representatives from previously uncharacterized protein families be selected preferentially?
- Should medically important proteins such as drug discovery targets be chosen first?
- Should there be efforts to exhaustively solve all structures within an individual organism? This is being considered for *Methanococcus jannaschii* and *Mycobacterium tuberculosis.*



Related Structures

Sequences producing significant alignments:		Score (bits)	E Value
gi|1827612|pdb|1OBP|A	Chain A, Odorant-Binding Protein From...	293	2e-80
gi|17943132|pdb|1HN2|A	Chain A, Crystal Structure Of Bovine...	292	4e-80
gi|21730166|pdb|1G85|A	Chain A, Crystal Structure Of Bovine...	291	5e-80
gi|2392495|pdb|1PBO|A	Chain A, Complex Of Bovine Odorant Bi...	291	8e-80
gi|12084609|pdb|1D2J|A	Chain A, Porcine Odorant Binding Pro...	115	5e-27
gi|4558087|pdb|1A3Y|B	Chain B, Odorant Binding Protein From...	111	1e-25
gi|4558086|pdb|1A3Y|A	Chain A, Odorant Binding Protein From...	110	2e-25
gi|5107505|pdb|1BJ7|	Bovine Lipocalin Allergen Bos D 2	77	2e-15
gi|5822446|pdb|2A2G|A	Chain A, The Crystal Structures Of A2...	72	1e-13
gi|15826041|pdb|1E5P|A	Chain A, Crystal Structure Of Aphrod...	70	5e-13
gi|7766863|pdb|1EW3|A	Chain A, Crystal Structure Of The Maj...	58	1e-09
gi|21465464|pdb|1GM6|A	Chain A, 3-D Structure Of A Salivary...	58	1e-09
gi|13399976|pdb|1I04|A	Chain A, Crystal Structure Of Mouse ...	56	7e-09
gi|494384|pdb|1MUP|	Major Urinary Protein Complex With 2-(...	55	9e-09
gi|8569601|pdb|1DF3|A	Chain A, Solution Structure Of A Reco...	55	1e-08
gi|4558223|pdb|2SHP|A	Chain A, Tyrosine Phosphatase Shp-2 >...	28	1.3
gi|640187|pdb|1AYA|A	Chain A, Tyrosine Phosphatase Syp (N-T...	28	1.8

FIGURE 9.12. *Structure entries can be retrieved from NCBI by performing a blastp search (with a protein query) or a blastx search (with DNA) restricting the output to the PDB database. Here, a search with bovine OBP (P07435) produces matches against a variety of lipocalins of known structure, including bovine lipocalin allergen Bos D 2, α2u-globulin, and the sex pheromone binding protein from female hamster, aphrodisin.*

- How can structures be solved for more proteins having transmembrane-spanning domains? These are among the most technically challenging proteins to study. Chang and Roth (2001) successfully solved the structure of a multidrug-resistant ABC transporter from *E. coli.* They screened 96,000 crystallization conditions to find several that were adequate for X-ray structure determination.

PROTEIN DATA BANK

Once a protein sequence is determined, there is one principal repository in which the structure is deposited: the Protein Data Bank (PDB) (Westbrook et al., 2002; Berman et al., 2002). A broad range of primary structural data is collected, such as atomic coordinates, chemical structures of cofactors, and descriptions of the crystal structure. The PDB then validates structures by assessing the quality of the deposited models and by how well they match experimental data.

The PDB was established at Brookhaven National Laboratories in Long Island in 1971 and moved to the Research Collaboratory for Structural Bioinformatics (RCSB) in 1998. The PDB is accessed at ▸http://www.rcsb.org/pdb/.

FIGURE 9.13. *A text search with the word "odorant" from the NCBI structure page (Fig. 9.11) results in a list of Entrez sequence entries for which structure data are available. An example from this list is the bovine OBP (PDB accession 1PBO). The result of clicking this link is shown in Figure 9.15.*

The main page of the PDB website is shown in Figure 9.5. This database currently has over 20,000 structure entries (Table 9.2), with new structures added at a rapid rate (Fig. 9.6). The database can be accessed directly using a PDB identifier, which is an accession number consisting of one number and three letters (e.g., 4HHB for hemoglobin). The PDB database can also be searched by keyword. The result of a keyword search for lipocalins is shown in Figure 9.7. By clicking the "Explore" feature for the bovine odorant-binding protein, one accesses the Structure Explorer feature (Fig. 9.8). This lists information such as the resolution, the space group, and the unit cell dimensions of the crystals. The "View Structure" option links to a series of tools to visualize the three-dimensional structure (Fig. 9.9). Some of the most common ways to access interactive tools are listed in Table 9.3. We will explore some of these in this chapter.

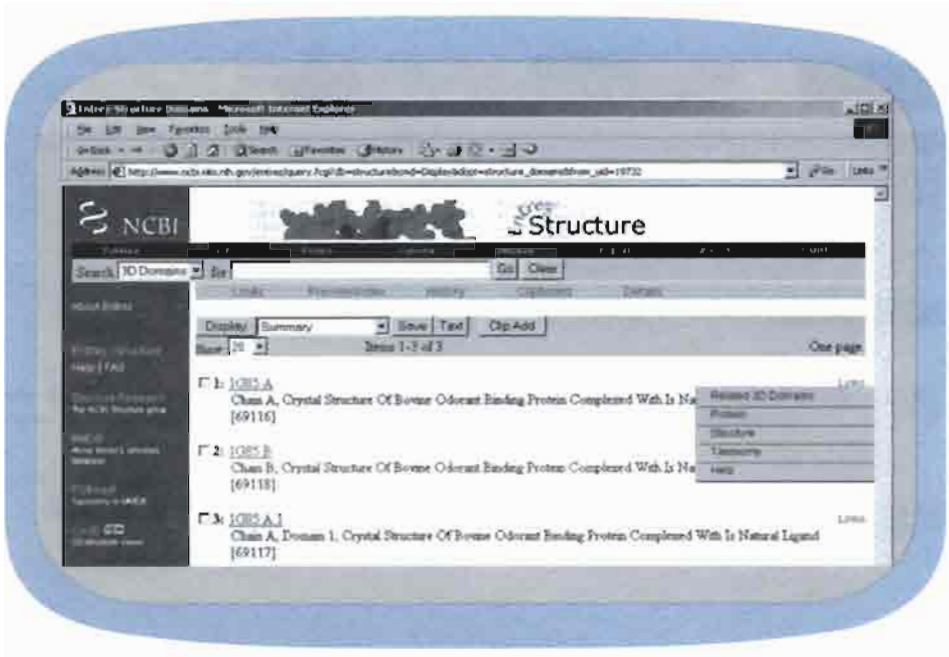


FIGURE 9.14. By clicking on the link "3D domains" (see Fig. 9.13), one obtains alternate PDB entries for a single protein structure, such as alternate protein chains. Further clicking the "Related 3D Domains" link here leads to several hundred structural database entries for lipocalins.

It is also possible to search within the PDB website using the FASTA program of Pearson and Lipman (1988). From the main page of PDB, select "SearchFields" (Fig. 9.5, arrow 1). This search page includes an option to perform a FASTA search, allowing convenient access to PDB structures related to a query. For short queries, the expect value should be raised using a command such as "-E 10."

In addition to the PDB, the European Bioinformatics Institute operates the Macromolecular Structure Database. This project is integrated with PDB and represents the European center for the collection of macromolecular structure data.

The Macromolecular Structure Database is at ▸http://www.ebi.ac.uk/msd/.

The PDB database occupies a central position in structural biology (Fig. 9.10). Several dozen other databases and web servers link directly to it or incorporate its data into their local resources (Web Resources, Table 9.6). We will consider these various resources in two ways as outlined in Figure 9.10: First we will explore NCBI and other sites that allow a single protein structure to be analyzed or several structures to be compared. Then we will explore databases such as SCOP, CATH, and Dali that create comprehensive classification systems or taxonomies for all protein structures.

Accessing PDB Entries at NCBI Website

There are three main ways to find a protein structure in the NCBI databases:

The NCBI structure page is at ▸http://www.ncbi.nlm.nih.gov/Structure/.

1. Keyword searches allow access to PDB structures. These keyword searches can be performed on the structure page (Fig. 9.11) or through Entrez.
2. One can search by protein similarity. To do this, use the Entrez protein database to select a protein of interest and look for a link to "Related Structures." Alternatively, perform a blastp search and restrict the output to the PDB database (Fig. 9.12).
3. One can search through the nucleotide database. It is also possible to use a blastx search with a DNA sequence as input, restricting the output to the PDB database.

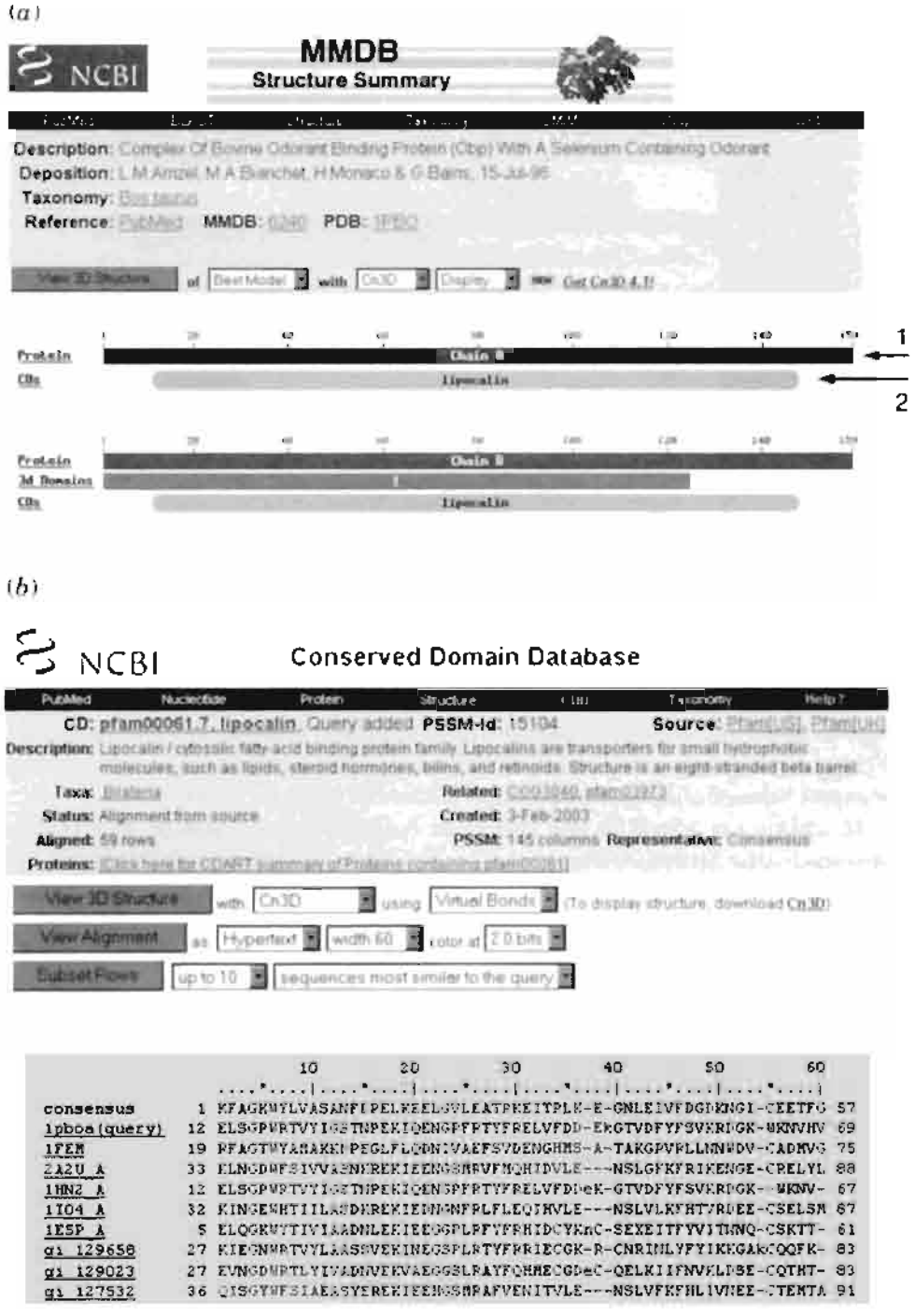


FIGURE 9.15. *(a) By clicking on the "1PBO" link from the list of Entrez protein structures (Figs. 9.13 and 9.14), one obtains this NCBI Molecular Modeling DataBase structure summary. By clicking on the "View 3D Structure" link, one can launch Cn3D or other structure viewers (RasMol, MAGE). There are links to "structure neighbors" (arrow 1) and to the Conserved Domain Database (CDD) entry (arrow 2). The CDD link provides a multiple sequence alignment, part of which is shown in (b).*

A search of Entrez structures for the odorant-binding protein yields a list of proteins with four-character PDB identifiers (Fig. 9.13). Clicking the "3D Domains" link shows the entries for different individual chains that were deposited in PDB for the same protein (Fig. 9.14). Clicking instead on the "1PBO" link shows the Molecular Modeling DataBase (MMDB) entry for this protein (Fig. 9.15). This is the main

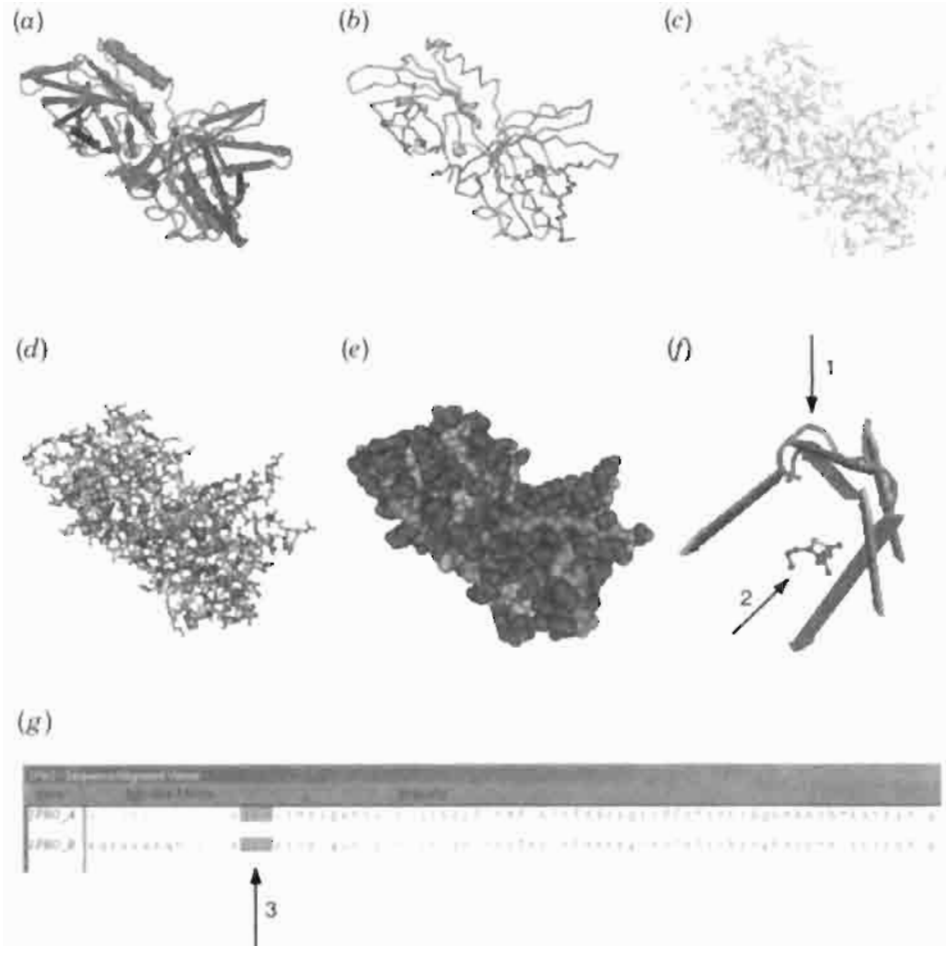


FIGURE 9.16. *The NCBI Cn3D viewer shows the secondary structure of the odorant-binding protein (1PBO), an eight-stranded beta barrel, in a variety of formats: (a) worms, (b) tubes, (c) wire, (d) ball and stick, and (e) space fill. (f) It is possible to rotate the image and zoom on any region of the protein. (g) The primary amino acid sequence is shown in a Sequence/Alignment Viewer. By clicking on a sequence consisting of the conserved GXW motif (arrow 3), it can be highlighted in the protein (arrow 1). There are a variety of other visualization options, such as an image of the ligand (an odorant) in the binding pocket of the protein (arrow 2).*

NCBI database entry for each protein structure (Wang, et al. 2000). It includes literature and taxonomy data, sequence neighbors (as defined by BLAST), structure neighbors (as defined by VAST; see below), and visualization options.

It is easy to view protein structures through MMDB by clicking the "View Structure" link. This requires that the Cn3D software be downloaded. Two windows open: a Cn3D Viewer and a OneD-Viewer (Fig. 9.16). The Cn3D Viewer shows the structure of the protein in seven different formats, and it can be rotated for exploration of the structure. The corresponding OneD-Viewer shows the amino acid sequence of the protein, including α helices and β sheets. Highlighting any individual amino acid residue or group of residues in either the Cn3D Viewer or the OneD-Viewer causes the corresponding region of the protein to be highlighted in the other viewer.

Cn3D is an acronym for "see in 3D."

In addition to investigating the structure of an individual protein, multiple protein structures can be compared simultaneously. Beginning at the main MMDB structure summary for a protein (Fig. 9.15), click "Structure Neighbors" to obtain a list of related proteins for which PDB entries are available (Fig. 9.17). This list is part of the Vector Alignment Search Tool (VAST). Selecting the entries for

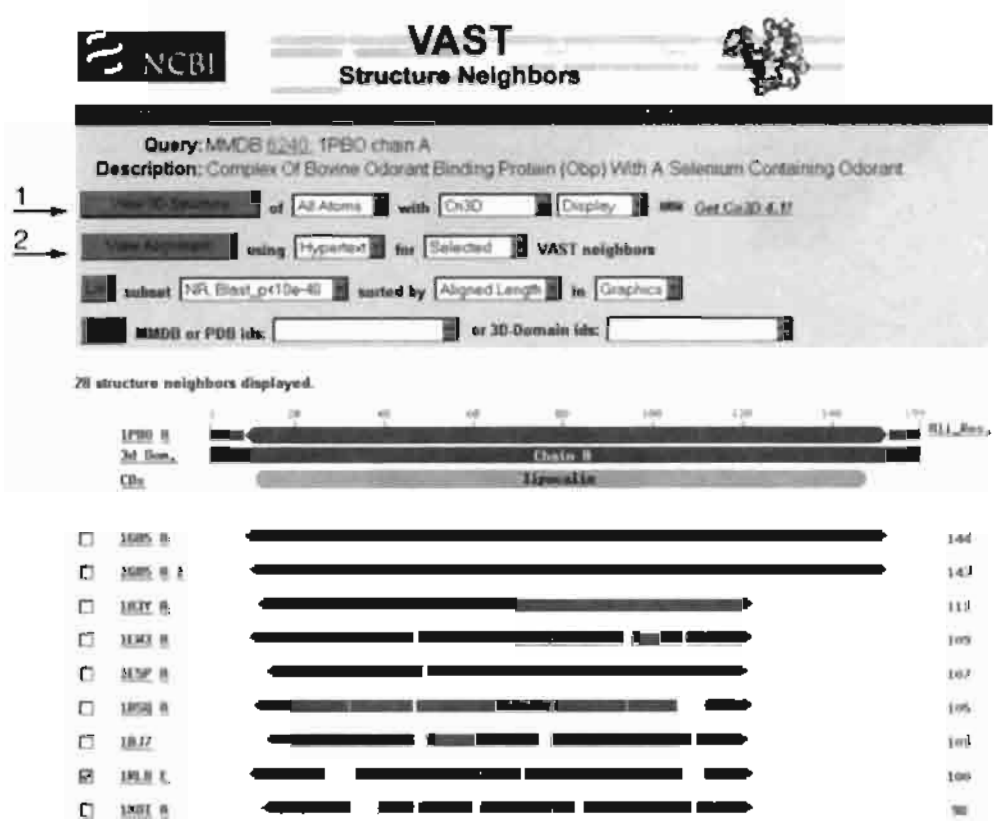


FIGURE 9.17. *By clicking the "Structure Neighbors" link (Fig. 9.15), one obtains the VAST (Vector Alignment Search Tool) page of related protein structures. Two or more structures may be compared with this tool via the Cn3D viewer (arrow 1), or a multiple sequence alignment may be viewed (arrow 2).*

retinol-binding protein (1RLB) and β-lactoglobulin (1BSO) in addition to the odorant-binding protein produces a Cn3D image of all three structures as well as a corresponding multiple sequence alignment in the two-dimensional viewer called DDV (Fig. 9.18). VAST provides many kinds of structural data (Box 9.2).

There are many other web-based interactive viewers available to visualize structures. The Swiss-PDB Viewer offers a large array of options, including Ramachandran plots, individualized coloring of any residue in the structure, and model-building algorithms. A view of the odorant-binding protein highlights the two chains of the homodimer in different colors (Fig. 9.19). This structure was solved by X-ray crystallography in the presence of an odorous ligand containing the heavy metal selenium. This ligand defines the binding pocket in each subunit (Fig. 9.19, arrows). This binding site is lined by hydrophobic residues that favor the binding of hydrophobic odorants.

Several structure visualization programs use Chime, a plug-in for web browsers that enables a variety of interactive viewing options. Examples are shown for a yeast *GAL4* protein complexed to DNA (Fig. 9.20) and myoglobin (Fig. 9.21). This software, like Swiss-PDB Viewer, allows a large number of viewing options.

Integrated Views of Universe of Protein Folds

The PDB database contains over 20,000 structures. We have examined how to view individual proteins and how to compare small numbers of structures. Several

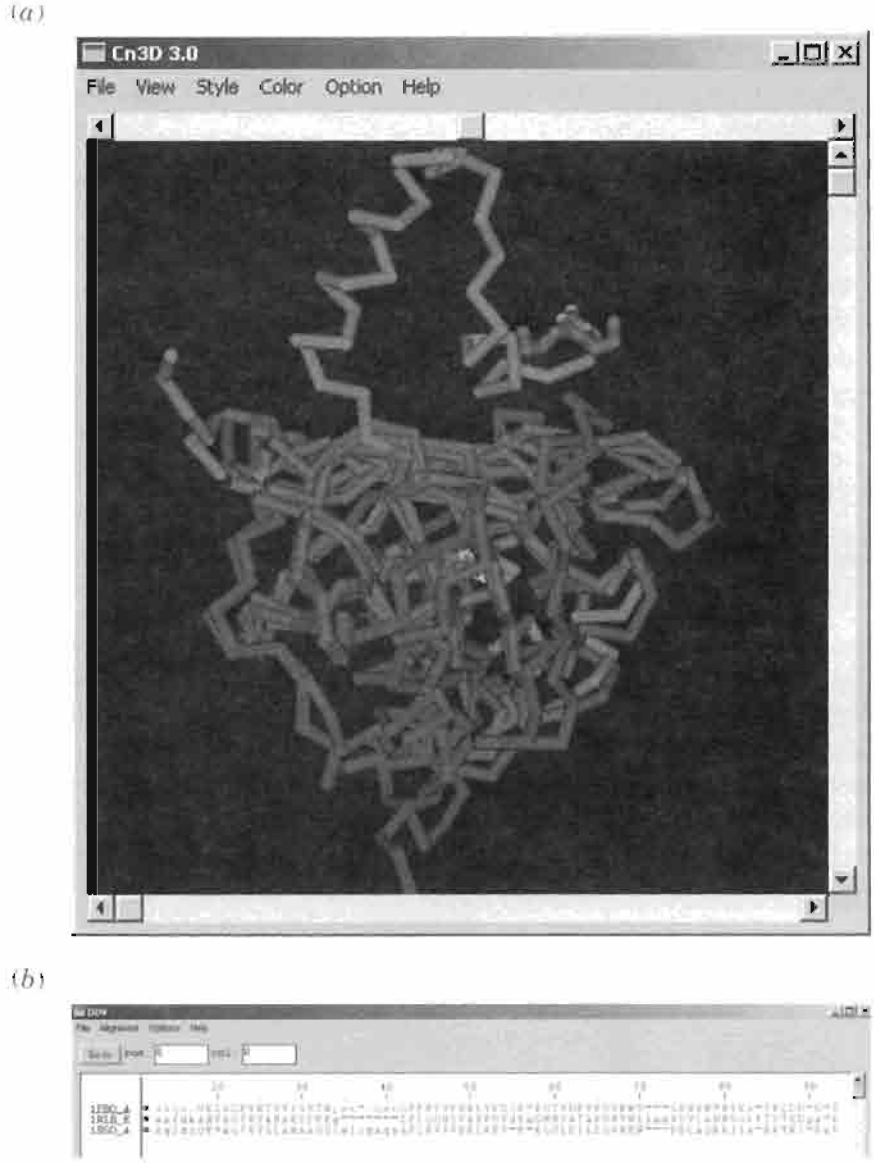


FIGURE 9.18. *Three structures that are selected in VAST (OBP, RBP, and β-lactoglobulin from Fig. 9.17) are compared (a) in the Cn3D viewer and (b) as a multiple sequence alignment in the two-dimension viewer DDV. Despite the relatively low sequence identity between these three proteins, they adopt highly similar three-dimensional folds.*

databases have been established to explore the broad question of the total protein fold space. How many different protein folds exist? How many structural groups are there? We will examine several of these databases: SCOP, CATH, Dali, and FSSP. These databases also permit searches for individual proteins.

Taxonomic System for Protein Structures: SCOP Database

The Structural Classification of Proteins (SCOP) database provides a comprehensive description of protein structures and evolutionary relationships based upon a hierarchical classification scheme (Fig. 9.22) (Murzin et al., 1995). At the top of the hierarchy are classes that are subsequently subdivided into folds, superfamilies, families, protein domains, and then individual PDB protein structure entries. The SCOP database can be navigated by browsing the hierarchy, by a keyword query or PDB identifier query, or by a homology search with a protein sequence.

The SCOP database is accessed at ▸http://scop.mrc-lmb.cam.ac.uk/scop/.

BOX 9-2
VAST Information

For each structural neighbor detected by VAST (such as Fig. 9.17), the following are listed in columns:

- Checkbox: The checkbox allows for selection of individual neighbors.
- PDB: The four-character PDB-Identifier of the structural neighbor.
- C: The PDB chain name.
- D: The MMDB domain identifier.
- RMSD: The root-mean-square superposition residual in angstroms. This is a descriptor of overall structural similarity
- NRES: The number of equivalent pairs of Cα atoms superimposed between the two structures. This number gives the alignment length, that is, how many residues have been used to calculate the three-dimensional superposition.
- %Id: Percent identical residues in the aligned sequence region.
- Description: A string parsed from PDB records

The main classes are listed in Table 9.4. The folds level of the hierarchy describes proteins sharing a particular secondary structure with the same arrangement and topology. However, different proteins with the same fold are not necessarily evolutionarily related.

As we continue down the SCOP hierarchy, we arrive at the level of the superfamily. Here proteins probably do share an evolutionary relationship, even if they share relatively low amino acid sequence identity in pairwise alignments. For example, the lipocalin superfamily in the SCOP database includes both the retinol-binding

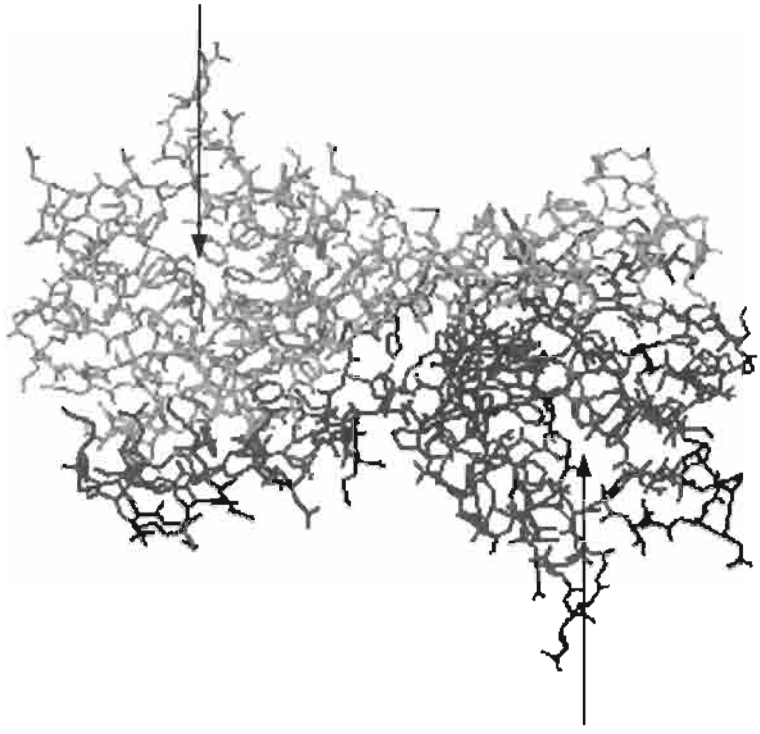


FIGURE 9.19. *The ExPASy analysis tools include the Swiss-PDB viewer (available at ▸http://www.expasy.org/spdbv/). This viewer offers far more options than Cn3D for more sophisticated analyses. Here, the two chains of the OBP dimer (1PBO) are visualized in two colors. This crystal structure was solved using a selenium odorant compound to identify the hydrophobic binding site on each subunit (arrows).*

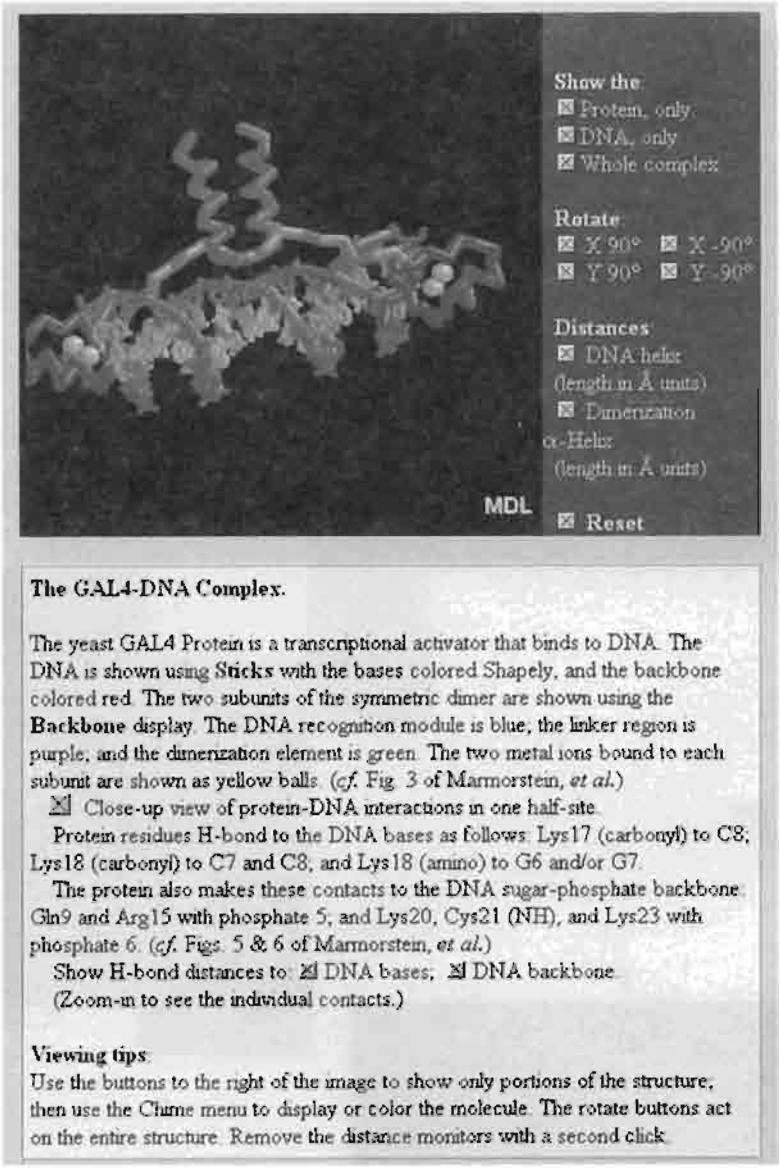


FIGURE 9.20. *Chime is a web browser plug-in that allows interactive viewing of PDB structures. This image displays the yeast GAL4 transcriptional activator bound to DNA (see Chapter 15). From* ▸*http://info.bio.cmu.edu/courses/03231//chime.tut/chime.html.*

protein family that we have used as an example and another group of carrier proteins exemplified by the fatty acid-binding proteins (FABPs). Like the lipocalins, most FABPs are small (15 kD), abundant, secreted proteins that bind hydrophobic ligands, and they generally have a glycine–X–tryptophan motif near the amino terminus of each protein. Pairwise sequence alignment fails to reveal significant matches, but the FABP family and the RBP lipocalin family are likely to be homologous. In the SCOP database, the lipocalin superfamily contains three groups: the RBP-like lipocalins, the FABPs, and a thrombin protein. (Some researchers have defined the calycin superfamily as consisting of RBP-like lipocalins, the FABPs, and the avidins (Flower et al., 2000).)

In the SCOP hierarchy members of a family have a clear evolutionary relationship. Usually, the structures of the proteins are related and the pairwise amino acid sequence identity is greater than 30%. In some cases, such as the lipocalins or the globins, some members of each family share as little as 15% identity, but the

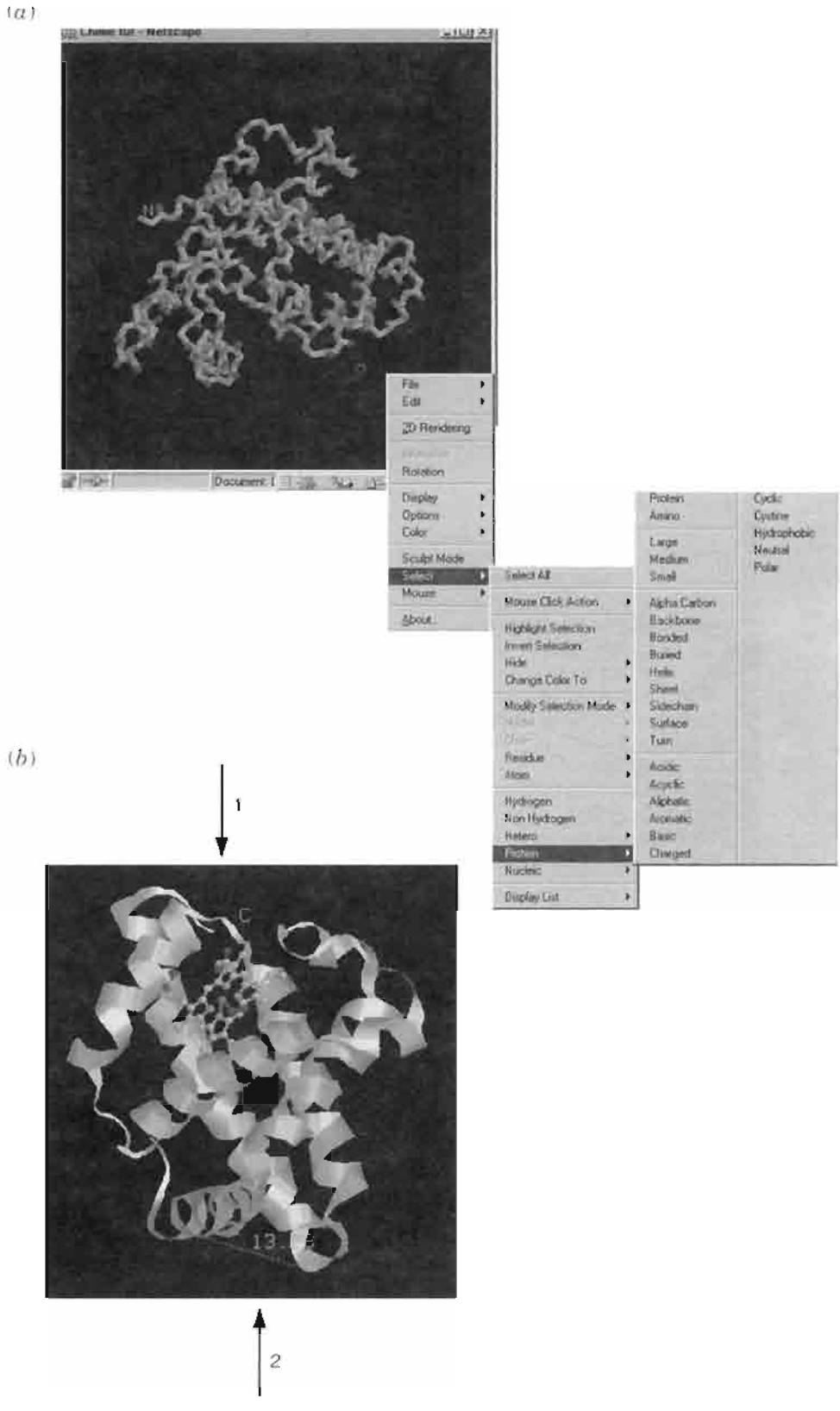


FIGURE 9.21. *(a) The Chime view of myoglobin includes many options for viewing the protein (pull-down menus). (b) For example, you can view the heme ligand (arrow 1) or specify the distance between any atoms (in angstroms; arrow 2).*

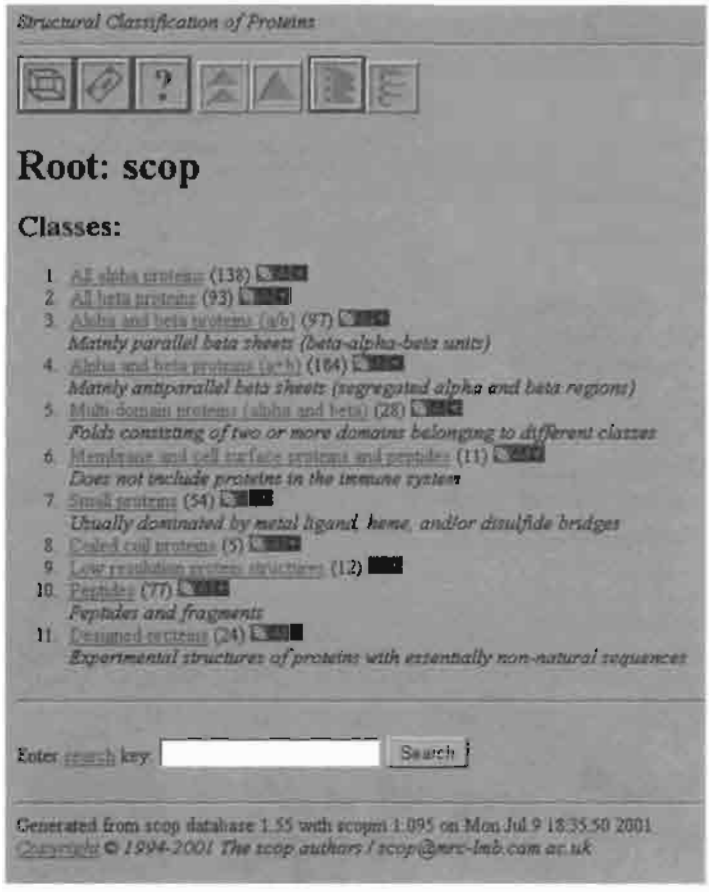


FIGURE 9.22. *The Structural Classification of Proteins (SCOP) database.*

assignment to the status of a family member is still unambiguous based upon structural and evolutionary considerations.

CATH Database

CATH is a hierarchical classification system that describes all known protein domain structures (Orengo et al., 1997; Pearl et al., 2000). CATH clusters proteins at four major levels: class (C), architecture (A), topology (T), and homologous superfamily (H).

CATH is accessed at ▸http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html.

At the highest level (class) the CATH database describes three main folds based on secondary-structure prediction: α, α and β, and β. Assignment at this level resembles the SCOP database system (Table 9.4). The architecture (A) level of CATH describes the shape of the domain structure as determined by the orientations of the secondary structures. Examples are the TIM barrel (named for triose phosphate isomerase) and jelly roll (Fig. 9.23). These assignments are made by expert judgment rather than by an automated process.

The topology (T) level of CATH describes fold families. Protein domains are clustered into families using the SSAP algorithm of Taylor and Orengo (1989a,b). Proteins sharing topologies in common are not necessarily homologous. In contrast, the homologous superfamily (H) level clusters proteins that are likely to share homology (i.e., descent from a common ancestor).

The SSAP algorithm compares two protein structures. It can be accessed at ▸http://www.biochem.ucl.ac.uk/cgi-bin/cath/GetSsapRasmol.pl.

TABLE 9-4 Release Notes from SCOP Database, Release 1.59 (October 2002)

For each fold, there are between one and dozens of superfamilies.

Class	Number of Folds	Number of Superfamilies	Number of Families	Notes[a]
All alpha proteins	151	252	393	–
All beta proteins	110	205	337	–
Alpha and beta proteins (α/β)	113	185	438	1
Alpha and beta proteins ($\alpha + \beta$)	208	295	454	2
Multidomain proteins	34	34	46	3
Membrane and cell surface proteins	12	19	31	4
Small proteins	58	83	128	5
Total	686	1073	1827	–

[a](1) Mainly parallel beta sheets (beta-alpha-beta units). (2) Mainly antiparallel beta sheets (segregated alpha and beta regions). (3) Folds consisting of two or more domains belonging to different classes. (4) Does not include proteins in the immune system. (5) Usually dominated by metal ligand, heme, and/or disulfide bridges.

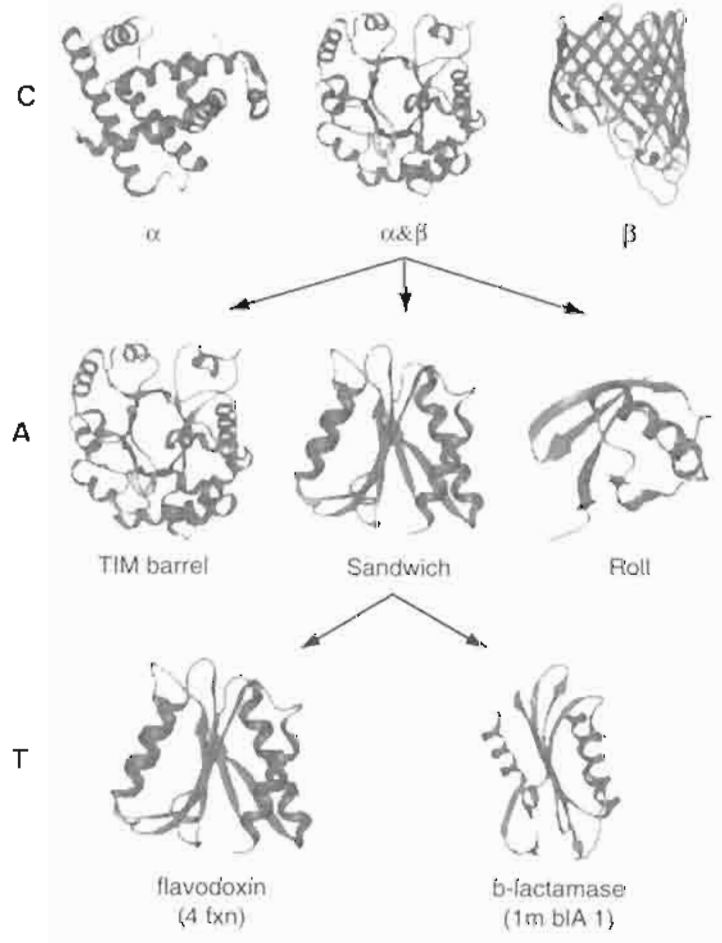


FIGURE 9.23. *The CATH resource organizes protein structures by a hierarchical scheme of class, architecture, topology (fold family), and homologous superfamily. From* ▸*http://www.biochem.ucl.ac.uk/bsm/cath_new/cath_info.html#C_Level.*

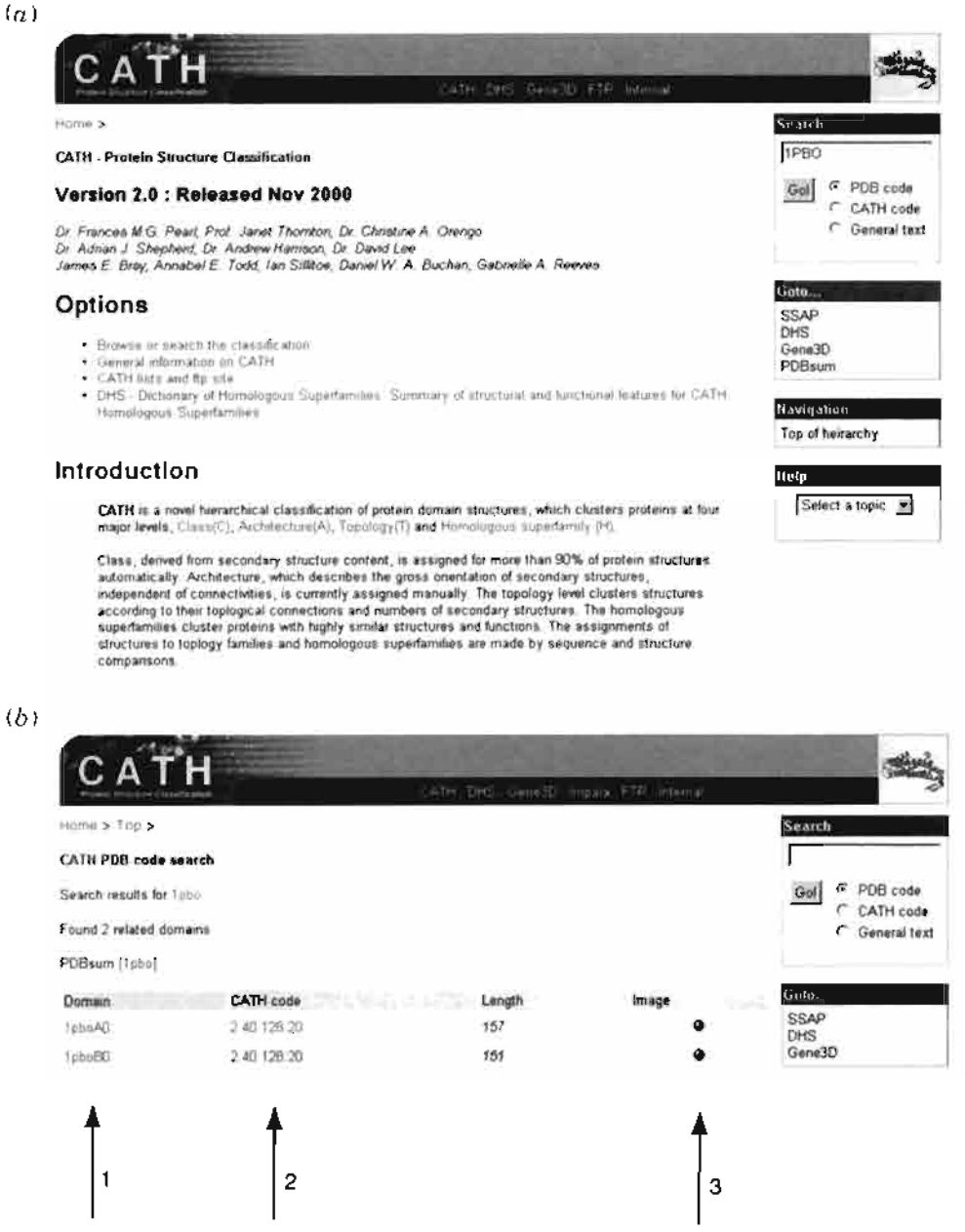


FIGURE 9.24. *(a) A search of the CATH database begins at the front page where a search can be performed by PDB code, CATH code, or general text. (b) A search for the PDB entry 1PBO yields the CATH result, including links to full-description pages (arrows 1 and 2) or a static image of the protein (arrow 3).*

A search of the CATH database demonstrates the usefulness of its hierarchical classification system. A query may be entered on the front page by PDB accession, CATH accession, or keyword (Fig. 9.24). This produces links to pages describing the odorant-binding protein, showing its description at each level of the CATH hierarchy (Fig. 9.25*a*). Links to additional lipocalins are also provided (Fig. 9.25*b*). A portion of the PDBsum link is shown in Figures 9.26 and 9.27. The PDBsum page shows links to several structure viewers, links to SwissProt and Pfam databases, and links to a large set of other databases that rely on PDB. Furthermore, one can explore the CATH Dictionary of Homologous Superfamilies (Fig. 9.28). In sum, the

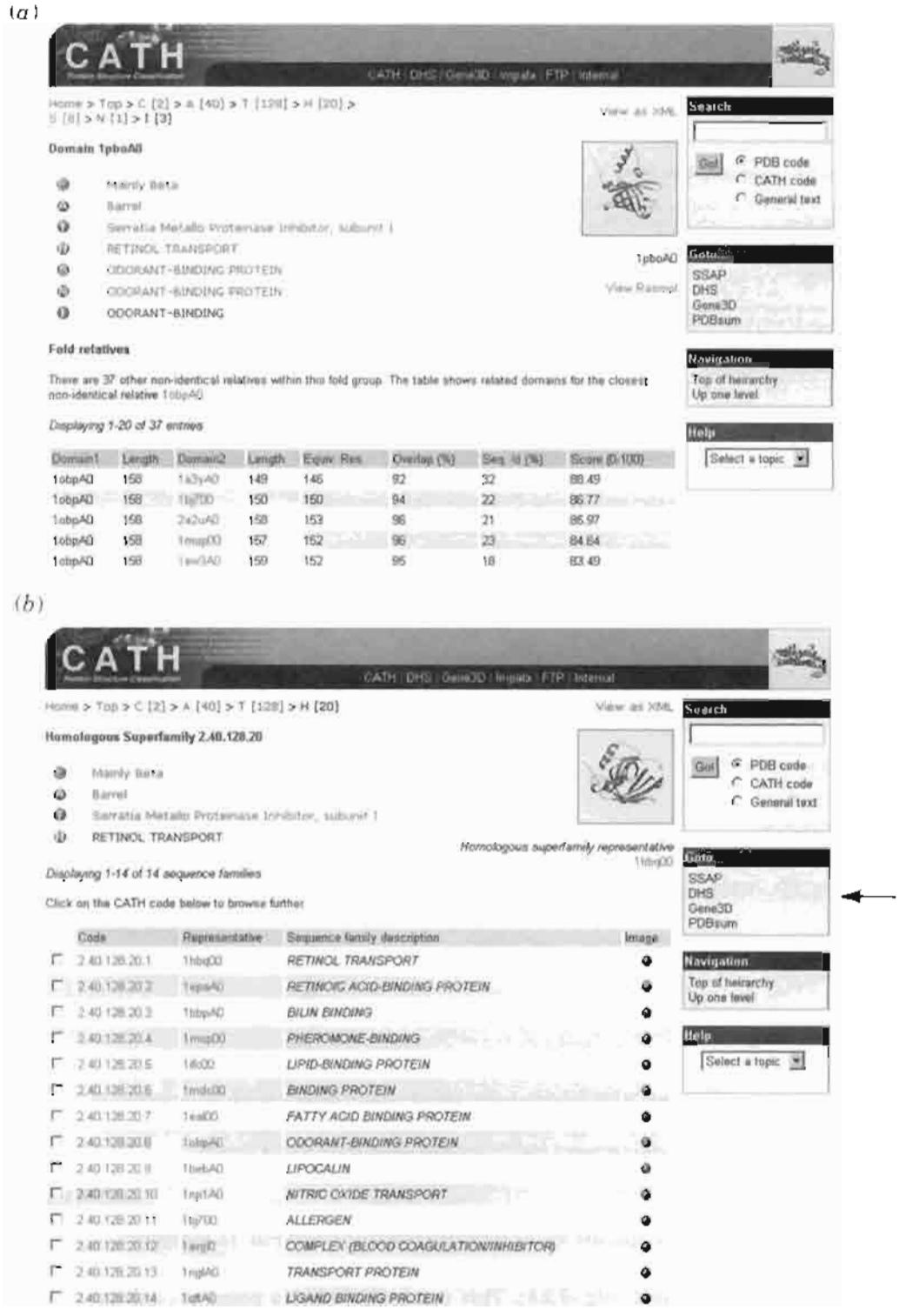


FIGURE 9.25. *A CATH database search (Fig. 9.24) results in outputs of (a) odorant-binding protein chains and (b) lipocalin neighbors. Additional links are available such as PDBSum (arrow 1).*

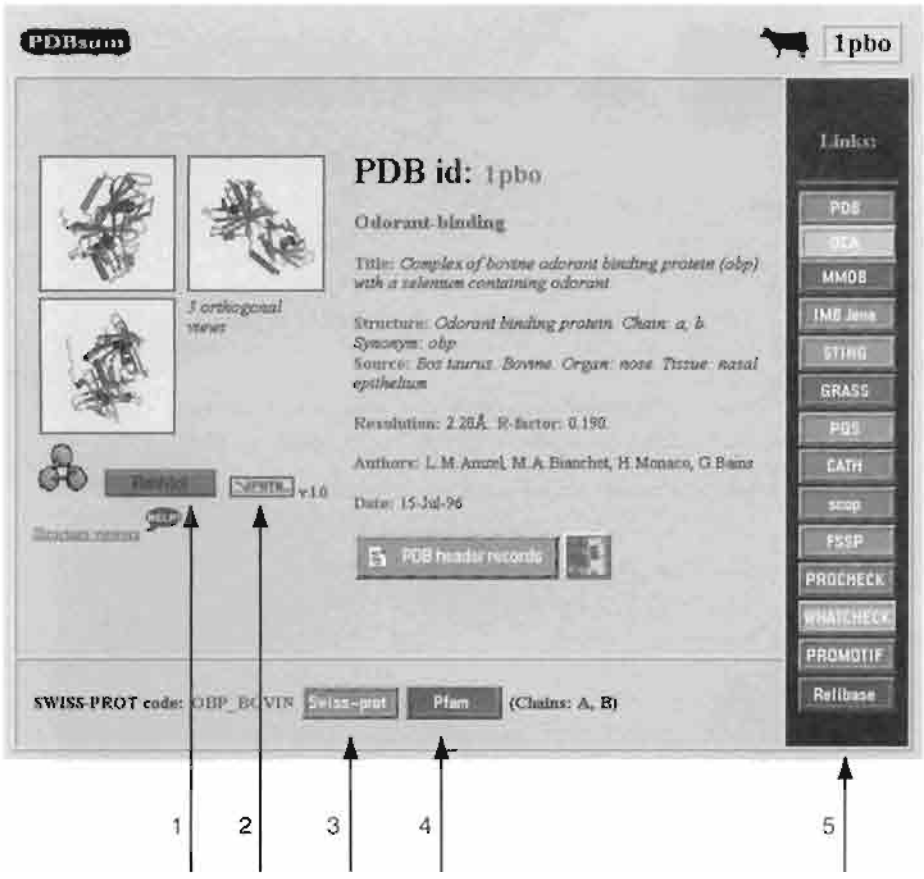


FIGURE 9.26. *The CATH database PDBsum entry for bovine OBP includes two tools to visualize the structure (RasMol, arrow 1, and VRML, arrow 2), links to SwissProt (arrow 3) and Pfam (arrow 4), and a variety of additional links to structural analyses of this protein (arrow 5). This PDBsum page contains further output (see Fig. 9.27).*

CATH database provides a deep and broad set of data on the structure of individual proteins, placing them in the context of a comprehensive taxonomy of protein structure.

Dali Domain Dictionary

The Dali Domain Dictionary contains a numerical taxonomy of all known structures in PDB (Dietmann et al., 2001). The classification scheme is based upon a comprehensive comparison of protein structures in PDB to create a map of fold space. A graph of structural similarity between all structures is generated and partitioned into clusters based upon several hierarchical levels:

- Supersecondary structural motifs (attractors in fold space)
- Topology of globular domains (fold types)
- Remote homologs (functional families)
- Homologs with sequence identity above 25% amino acid identity (sequence families)

In September 2000 the Dali classification scheme contained over 10,000 PDB entries comprising 17,100 chains (Dietmann et al., 2001). These were divided into five attractor regions, over 1300 fold types, 2500 functional families, and 3700 domain sequence families. Dali is at ▸ http://www2.ebi.ac.uk/dali/.

Each protein domain in Dali is assigned a domain classification number. These are searchable by keyword. The main query page is shown in Figure 9.29, and part of the entry for the odorant-binding protein (DC.6.71.1.5) is shown in Figure 9.30. Notably, this database integrates a variety of sequence and structural information

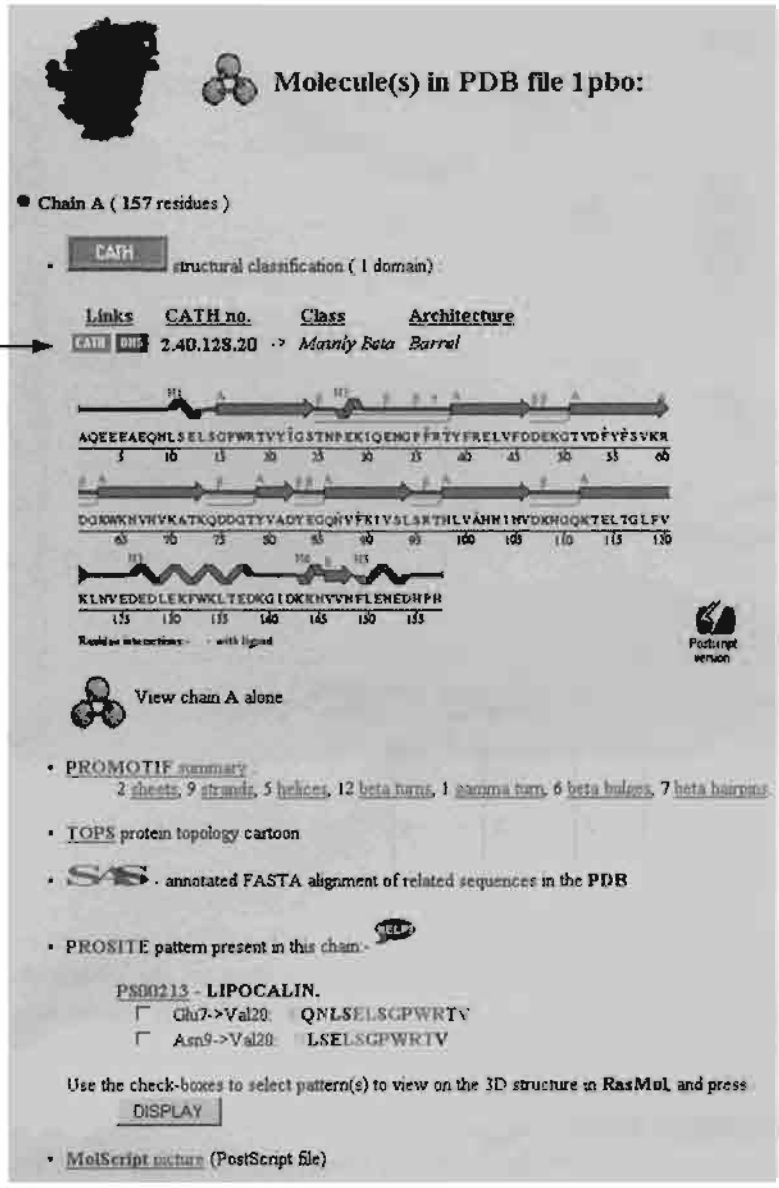


FIGURE 9.27. *Continuation of the output page for PDBsum in the CATH database (refer to Fig. 9.26). This entry shows the amino acid sequence of bovine OBP with a cartoon of the β sheets (solid arrows) and α helices determined from the crystal structure. Many additional links are provided, including to the CATH scheme and to the Dictionary of Homologous Superfamilies (arrow 1).*

for entries within domain classes, such as secondary-structure predictions and solvent accessibility.

FSSP Database

FSSP is at ▸http://www.ebi.ac.uk/dali/fssp.

The fold classification based on structure-structure alignment of proteins (FSSP) database is based upon a comprehensive comparison of all structures in PDB against each other (Fig. 9.31) (Holm and Sander, 1996a). The FSSP database stores all PDB structures greater than 30 amino acids in length. PDB structures are divided into "representative sets" which exclude all sequence homologs (i.e., protein pairs sharing greater than 25% amino acid identity). A fold classification of the representative set is accessed via a fold tree table (Fig. 9.32). Clicking on the entry for odorant-binding protein (Fig. 9.32, arrow 1) yields a set of structural neighbors (Fig. 9.33). A subset of these can be selected for a multiple sequence alignment (Fig. 9.34) and for viewing a superimposition of their three-dimensional structures.

Dictionary of Homologous Superfamilies - 2.40.128.20 -

CATH Code: 2.40.128.20

Class	2	Mainly Beta
Architecture	40	Barrel
Topology	128	Serratia Metallo Proteinase Inhibitor, subunit 1
Homologous superfamily	20	1cuoA0

Summary Information

No. of N-level Representatives	34
No. of Pairs	561
No. of Proteins in Multiple Structural Alignment (corelist)	34 (4)
No. of Positions in Multiple Structural Alignment	311
Sequence Identity Cutoff used to Generate Alignment	25
Structural Similarity (SSAP) Score used to Generate Alignment	80
No. of Prosite Entries	33
No. of Enzyme Classification Entries	0
No. of SWISS-PROT Entries	34
No. of Ligand Entries	26
No. of Alignment Positions with Ligand Interactions	77
Total No. of Ligand Interacting Residues (All Proteins)	238

FIGURE 9.28. *The CATH entry for bovine OBP includes a Dictionary of Homologous Superfamilies (DHS) link (Fig. 9.27, arrow 1) leading to this page. In addition to the CATH code and summary information shown here, the page also describes protein domains in the multiple structural alignment, an alignment download in several different formats, a list of Prosite patterns in the alignment, enzyme classification numbers (there are none in this example), SwissProt descriptions and keywords, protein-ligand binding patterns in the alignment, and a matrix of position-specific ligand information.*

COMPUTATIONAL BIOLOGY APPROACHES TO STRUCTURE

Comparative Modeling

While approximately 20,000 protein structures have been deposited in PDB, approximately 1,000,000 protein sequences have been deposited in the SwissProt/TrEMBL databases (June 2003). For the vast majority of proteins, the assignment

Dali Domain Dictionary v.3

Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L (2001) Nucleic Acids Res 29, 55-57.

- Dali Classification - start interactive browsing
- Dali Domain Definitions - download computer-readable database (tab-delimited table)
- Search PDB codes (example: 1ppt) or protein names (example: tyrosine kinase)

lipocalin submit reset

FIGURE 9.29. *The Dali Domain Dictionary (http://www.ebi.ac.uk/dali/domain) is a numerical taxonomy of protein structures in the PDB. The taxonomy is based upon structural, functional, and sequence similarities between proteins (Dietmann et al., 2001). It is based upon a map of "fold space" derived from a comprehensive comparison of all protein structures in PDB.*

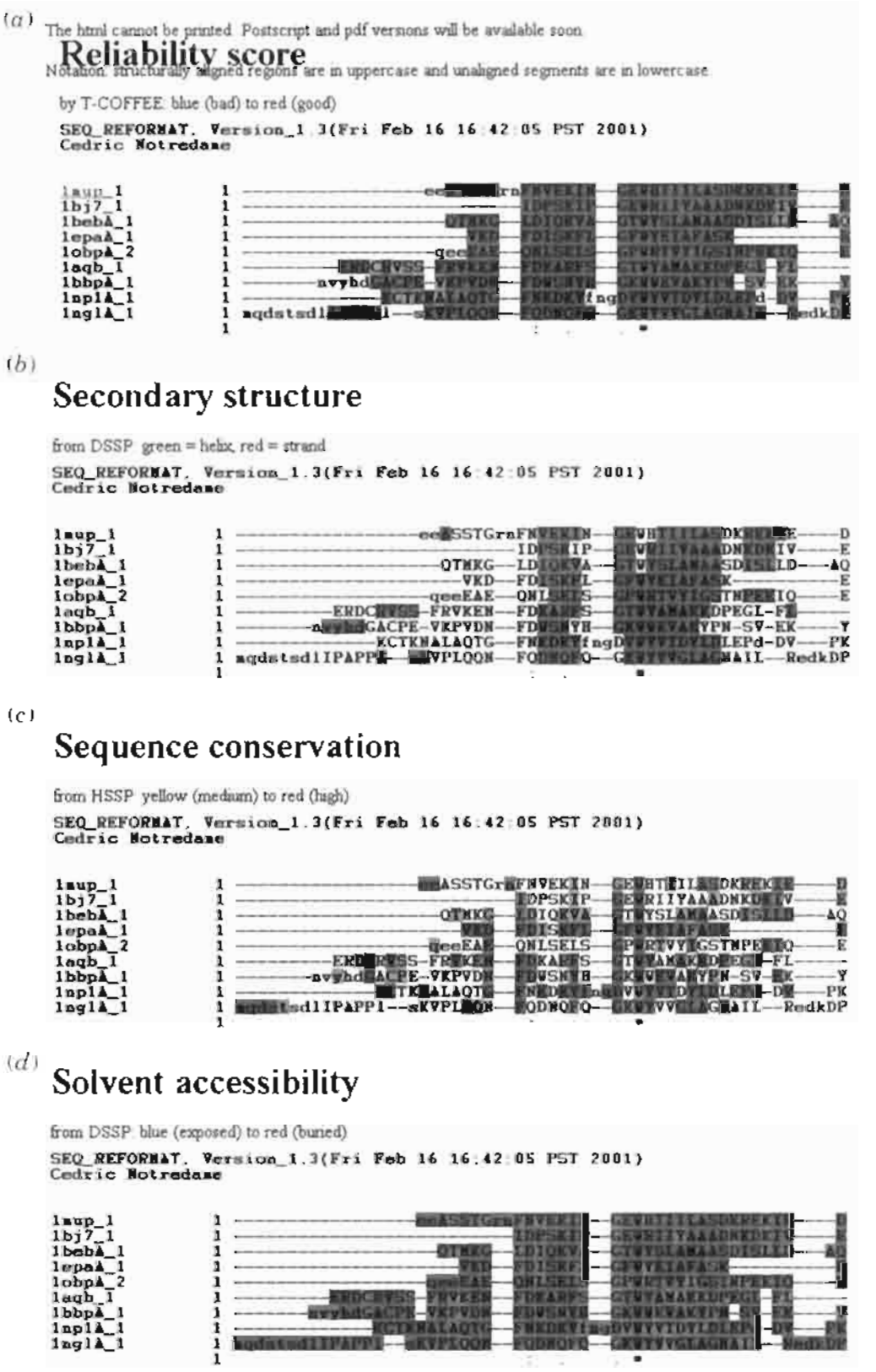
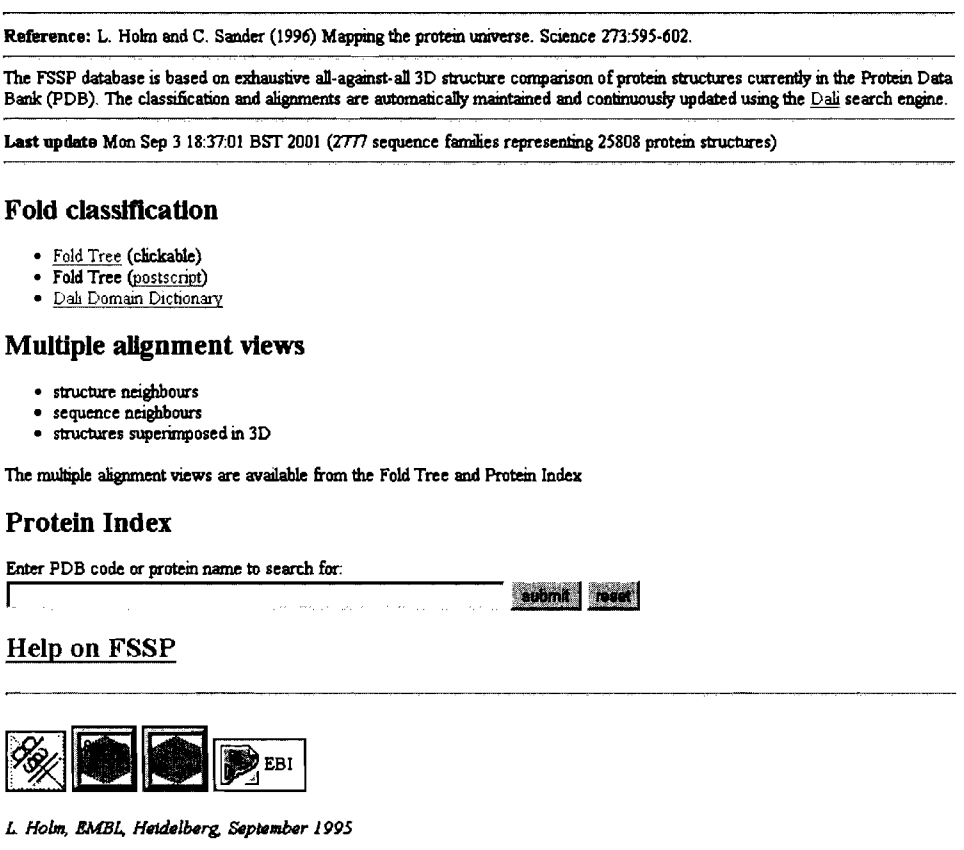


FIGURE 9.30. The Dali server provides information on multiple sequence alignments based on structural information, including, (a) reliability, (b) secondary structure, (c) sequence conservation, and (d) solvent accessibility.



Fold classification based on Structure-Structure alignment of Proteins (FSSP)

Reference: L. Holm and C. Sander (1996) Mapping the protein universe. Science 273:595-602.

The FSSP database is based on exhaustive all-against-all 3D structure comparison of protein structures currently in the Protein Data Bank (PDB). The classification and alignments are automatically maintained and continuously updated using the Dali search engine.

Last update Mon Sep 3 18:37:01 BST 2001 (2777 sequence families representing 25808 protein structures)

Fold classification

- Fold Tree (clickable)
- Fold Tree (postscript)
- Dali Domain Dictionary

Multiple alignment views

- structure neighbours
- sequence neighbours
- structures superimposed in 3D

The multiple alignment views are available from the Fold Tree and Protein Index

Protein Index

Enter PDB code or protein name to search for:

submit reset

Help on FSSP

EBI

L. Holm, EMBL, Heidelberg, September 1995

FIGURE 9.31. *The FSSP database is at the European Bioinformatics Institute (http://www.ebi.ac.uk:80/dali/fssp/) (Holm and Sander, 1996a).*

of structural models relies on computational biology approaches rather than experimental determination. As protein structures continue to be solved by X-ray crystallography and NMR spectroscopy, the most reliable method of modeling and evaluating new structures is by comparison to previously known structures (Jones, 2001; Baker and Sali, 2001). This is the method of comparative modeling of protein structure, also called homology modeling (Fig. 9.35). This method is fundamental to the field of structural genomics.

Comparative modeling consists of four sequential steps (Marti-Renom et al., 2000).

1. Fold assignment and template selection are performed. This can be accomplished by searching for homologous protein sequences and/or structures with tools such as BLAST and PSI-BLAST. The target can be queried against databases described in this chapter, such as PDB, CATH, and SCOP. As part of this analysis, structurally conserved regions and structurally variable regions are identified. It is common for structurally variable regions to correspond to loops and turns, often at the exterior of a protein.
2. The target is aligned with the template. As for any alignment problem, it is especially difficult to determine accurate alignments for distantly related proteins. For 30% sequence identity between a target and a template protein, the two proteins are likely to have a similar structure if the length of

255.1.1.1.1.1	1f3uA	alignment	transcription initiation factor iif, beta subunit fragment
255.1.2.1.1.1	1f3uB	alignment	transcription initiation factor iif, beta subunit fragment
256.1.1.1.1.1	1yua	alignment	topoisomerase i fragment
257.1.1.1.1.1	2omf	alignment	matrix porin outer membrane protein f (matrix porin, ompf p
257.1.1.1.2.1	1e54A	alignment	outer membrane porin protein 32 (omp32) omp32
257.1.1.1.3.1	2prn	alignment	porin Mutant biological_unit
257.1.1.1.4.1	2por	alignment	Porin (crystal form b)
257.1.1.2.1.1	2mprA	alignment	maltoporin fragment (lam-b, mal-1) biological_unit
257.1.1.2.1.2	1a0tP	alignment	sucrose-specific porin biological_unit
257.1.1.3.1.1	1fepA	alignment	ferric enterobactin receptor (fepa)
257.1.1.3.2.1	1by5A	alignment	ferric hydroxamate uptake protein (fhua) ferrichrome
257.1.2.1.1.1	1qd5A	alignment	outer membrane phospholipase a Mutant
257.1.3.1.1.1	4bcl	alignment	bacteriochlorophyll a protein biological_unit
257.1.4.1.1.1	1qj8A	alignment	outer membrane protein x Mutant
257.1.4.2.1.1	1bxwA	alignment	outer membrane protein a fragment Mutant
257.1.5.1.1.1	1swuA	alignment	streptavidin Mutant
257.2.1.1.1.1	1dwnA	alignment	phage coat protein
257.2.1.2.1.1	1qbeA	alignment	bacteriophage q beta capsid
257.2.1.3.1.1	1mse	alignment	Bacteriophage ms2 unassembled coat protein dimer (translati
258.1.1.1.1.1	1bj7	alignment	d 2 (dander major allergen bda20, dermal allergen bda20) bi
258.1.1.1.2.1	1mup	alignment	Major urinary protein complex with 2-(sec-butyl) thiazoline
258.1.1.1.3.1	1bebA	alignment	beta-lactoglobulin biological_unit
258.1.1.1.4.1	1epaA	alignment	Epididymal retinoic acid-binding protein (androgen dependen
258.1.1.2.1.1	1bbpA	alignment	Bilin binding protein (BBP)
258.1.1.2.2.1	1aqb	alignment	retinol-binding protein (rbp)
258.1.1.3.1.1	1np1A	alignment	nitrophorin 1 (np1) biological_unit
258.1.1.4.1.1	1ngla	alignment	ngal fragment (hngal, hnl) biological_unit
258.1.1.5.1.1	1qfta	alignment	female-specific histamine binding protein 2 female-specific
258.1.1.6.1.1	1obpA	alignment	odorant-binding protein (obp)
258.1.2.1.1.1	1hmt	alignment	Fatty acid binding protein (human muscle, m-fabp) complexed
258.1.3.1.1.1	1avgI	alignment	thrombin triabin
258.2.1.1.1.1	1smpI	alignment	serratia metallo proteinase erwinia chrysanthemi inhibitor
259.1.1.1.1.1	1bvqA	alignment	4-hydroxybenzoyl coa thioesterase
259.1.1.2.1.1	1c8uA	alignment	acyl-coa thioesterase ii
259.1.1.3.1.1	1mkaA	alignment	beta-hydroxydecanoyl thiol ester dehydrase (beta-hydroxydec
260.1.1.1.1.1	1ffkU	alignment	23s rrna 5s rrna ribosomal protein l2 (50s ribosomal protei
261.1.1.1.1.1	1bvl	alignment	bet v 1 (major pollen allergen bet v 1-a)
261.1.1.2.1.1	1em2A	alignment	mln64 protein fragment Mutant
261.1.1.3.1.1	1fvsA	alignment	phosphatidylinositol transfer protein (pitp)
261.1.2.1.1.1	3pmgA	alignment	alpha-d-glucose-1,6-bisphosphate (phosphoglucomutase)
261.2.1.1.1.1	1jenA	alignment	s-adenosylmethionine decarboxylase (beta chain) (adometdc,
261.3.1.1.1.1	1d8iA	alignment	mrna triphosphatase cet1
261.4.1.1.1.1	1eaf	alignment	Dihydrolipoyl transacetylase (catalytic domain (residues 38
261.4.1.1.2.1	3cla	alignment	Type III chloramphenicol acetyltransferase (CAT-III=) compl
262.1.1.1.1.1	1fmaE	alignment	molybdopterin convertin factor, subunit 1 molybdopterin con
262.2.1.1.1.1	2fua	alignment	l-fuculose-1-phosphate aldolase
263.1.1.1.1.1	1g8jB	alignment	arsenite oxidase fragment arsenite oxidase fragment
263.1.1.2.1.1	1fqtA	alignment	rieske-type ferredoxin of biphenyl dioxygenase (biphenyl di
263.1.1.3.1.1	1rie	alignment	rieske iron-sulfur protein fragment (ubiquinol-cytochrome c
263.1.1.4.1.1	1rfs	alignment	rieske protein fragment (rieske iron-sulfur protein, risp)

1

FIGURE 9.32. *The FSSP fold tree is a table of protein folds labeled with identifiers. The table can be searched by name or identifier. The position of the OBP is indicated (arrow 1).*

the aligned region is sufficient (e.g., more than 60 amino acids). The use of multiple sequence alignments (Chapter 10) can be especially useful.

3. A model is built. A variety of approaches are employed, such as rigid-body assembly and segment matching.
4. The model must be evaluated.

There are several principal types of errors that occur in comparative modeling (see Marti-Renom et al., 2000):

- Errors in side-chain packing
- Distortions within correctly aligned regions
- Errors in regions of a target that lack a match to a template
- Errors in sequence alignment
- Use of incorrect templates

The accuracy of protein structure prediction is closely related to the percent sequence identity between a target protein and its template (Fig. 9.36). When the two proteins share 50% amino acid identity or more, the quality of the model is usually excellent. For example, the root-mean-square deviation (RMSD) for the

FSSP: select structural neighbours of 1obpA

Please cite: L. Holm and C. Sander (1996) Science 273(5275):595-60

Select (check) structural neighbours to display

3D superimposition | Multiple alignment (wide) | Multiple alignment (narrow) | Multiple families (wide)
Multiple families (narrow) | Reset selection

	STRID2	Z	RMSD	LALI	LSEQ2	%IDE	PROTEIN
☐	1obpA	27.2	0.0	158	158	100	odorant-binding protein (obp)
☐	1pboB	23.0	0.3	145	151	99	odorant binding protein (obp)
☐	1obpB	22.0	0.8	145	155	100	odorant-binding protein (obp)
☐	1pboA	21.7	0.8	145	157	99	odorant binding protein (obp)
☐	1dzpB	16.7	1.8	115	147	42	odorant-binding protein (pig obp)
☐	1a3yB	16.7	1.8	115	147	42	odorant binding protein biological_unit
☐	1dzmB	16.5	1.8	115	147	42	odorant-binding protein (pig obp)
☐	1dzjA	16.5	1.9	115	148	42	odorant-binding protein (pig obp)
☐	1e06A	16.4	1.9	115	149	42	odorant-binding protein synonym pig obp
☐	1a3yA	16.4	1.8	115	149	42	odorant binding protein biological_unit
☐	1dzpA	16.4	1.9	115	149	42	odorant-binding protein (pig obp)
☐	1dzmA	16.3	1.9	115	149	42	odorant-binding protein (pig obp)
☐	1e00A	16.3	1.9	115	149	42	odorant-binding protein synonym pig obp
☐	1dzjB	15.3	1.8	115	148	42	odorant-binding protein (pig obp)
☐	1dzkB	15.3	1.8	115	148	42	odorant-binding protein (pig obp)
☐	1e06B	15.3	1.8	115	147	42	odorant-binding protein synonym pig obp
☐	1e00B	15.2	1.8	115	148	42	odorant-binding protein synonym pig obp
☐	1dzkA	15.1	1.9	115	148	42	odorant-binding protein (pig obp)
☐	1e02B	15.1	2.0	116	147	41	odorant-binding protein synonym pig obp
☐	1hqpA	15.0	1.8	115	149	42	odorant-binding protein
☐	1bj7	15.0	2.2	117	150	26	d 2 (dander major allergen bda20, dermal allergen bda20
☐	1e02A	14.9	1.9	115	149	42	odorant-binding protein synonym pig obp
☐	1mup	13.8	2.4	118	157	28	Major urinary protein complex with 2-(sec-butyl) thiazo
☐	1bebA	12.0	2.3	116	156	14	beta-lactoglobulin biological_unit
☐	1epaA	8.7	2.8	104	151	12	Epididymal retinoic acid-binding protein (androgen depe
☐	1aqb	7.9	3.0	109	175	14	retinol-binding protein (rbp)
☐	1npiA	7.4	3.3	111	184	8	nitrophorin 1 (npi) biological_unit
☐	1ngla	7.4	4.1	110	179	11	ngal fragment (hngal, hnl) biological_unit
☐	1bbpA	6.7	3.8	111	173	13	Bilin binding protein (BBP)
☐	1gftA	5.8	3.4	107	175	10	female-specific histamine binding protein 2 female-spec
☐	1bxwA	5.4	3.7	101	172	3	outer membrane protein a fragment Mutant

FIGURE 9.33. *The FSSP tool allows structural neighbors of a protein to be selected for further analysis and comparison.*

FSSP: structural neighbours of 1obpA

Please cite: L. Holm and C. Sander (1996) Science 273(5275):595-60.

Structural alignment by Dali

Notation: Uppercase: structurally equivalent with 1obpA; lowercase: structurally non-equivalent with 1obpA

```
1obpA     QEEE........AEQNLSELSGPWRTVYIGSTNPEKIQE.NGPFRT.YFRELVFDDEKGT
1obpA     QEEE........AEQNLSELSGPWRTVYIGSTNPEKIQE.NGPFRT.YFRELVFDDEKGT
1bebA     .........qtmKGLDIQKVAGTWYSLAMAASDISLLDAqSAPLRV.YVEELKPT.PEGD
1aqb      erdcrvssfrvkENFDKARFSGTWYAMAKKDPEG.......LFLQdnIVAEFSVD.ENGH

1obpA     VDFYFSVKRD.....GKWKNVHVKATKQDDG.TYVADY........EGQNVFKIVSLS.R
1obpA     VDFYFSVKRD.....GKWKNVHVKATKQDDG.TYVADY........EGQNVFKIVSLS.R
1bebA     LEILLQKWEN.....GECAQKKIIAEKTKIPaVFKIDA........LNENKVLVLDTDyK
1aqb      MSATAKGRVRllnnwDVCADMVGTFTDTEDPaKFKMKYwgvasflqKGNDDHWIIDTDyD

1obpA     THLVAHNINVDKHGQ..TTELTGLFVKLNVEDEDLEKFWKLTEDKGIDKKNVVNFLENED
1obpA     THLVAHNINVDKHGQ..TTELTGLFVKLNVEDEDLEKFWKLTEDKGIDKKNVVNFLENED
1bebA     KYLLFCMENS.aepE..QSLVCQCLVRT.pevddealekfdkalkalpmhirlsfnptql
1aqb      TYAVQYSCRLqnldgtcADSYSFVFARD..phgfspevqkivrqrqeelclarqyriith

1obpA     HPHPE
1obpA     HPHPE
1bebA     EEQC.
1aqb      NGYCD
```

return to FSSP home page / Dali Domain Dictionary

(C) L. Holm, EMBL-EBI, Hinxton, May 1996

FIGURE 9.34. *FSSP structural neighbors of OBP (1obpA), including β-lactoglobulin and retinol-binding protein, are multiply aligned.*

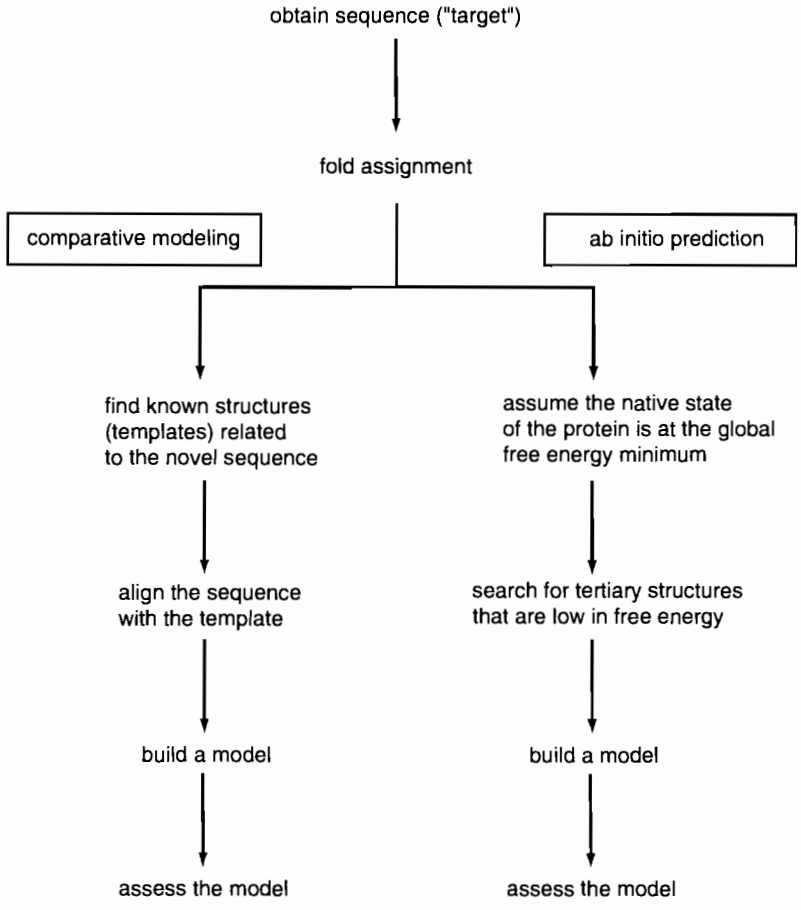


FIGURE 9.35. *Approaches to predicting protein structures (Baker and Sali, 2001). Comparative modeling is the most powerful approach when a target sequence has any indications of homology with a known structure. In the absence of homologous structures, ab initio prediction is used to model protein structure.*

FIGURE 9.36. *Protein structure prediction and accuracy as a function of the relatedness of a novel structure to a known template. Modified from Baker and Sali (2001). Abbreviation: aa, amino acids. Used with permission.*

FIGURE 9.37. *ExPASy SWISS-MODEL is a comparative modeling server that is useful to predict the structure of a protein. See* ▸*http://www.expasy.ch/swissmod/SWISS-MODEL.html.*

main-chain atoms tends to be 1 Å in such cases. Model accuracy declines when comparative models rely on 30–50% identity, and the error rate rises rapidly below 30% identity. De novo models are able to generate low-resolution structure models.

Many web servers offer comparative modeling, such as SWISS-MODEL at ExPASy (Fig. 9.37), the Predict Protein server at Columbia (Figs. 9.38 and 9.39), WHAT IF at CMBI (Netherlands), and the San Diego Protein Structure Homology Modeling Server. Others are listed in Table 9.6 (under Web Resources).

Ab Initio Prediction

In the absence of detectable homologs, protein structure may be assessed by ab initio (or de novo) structure prediction (Fig. 9.35). Here, protein folding is modeled based on global free-energy minimum estimates, and there is no overall comparison to known structures (Osguthorpe, 2000; Simons et al., 2001). While the resolution of ab initio methods is generally low, this approach is useful to provide structural models.

Bystroff and Shao (2002) incorporated the Rosetta Stone and related ab initio methods into a web-accessible server for protein structure prediction at ▸http://www.bioinfo.rpi.edu/~bystrc/hmmstr/server.html.

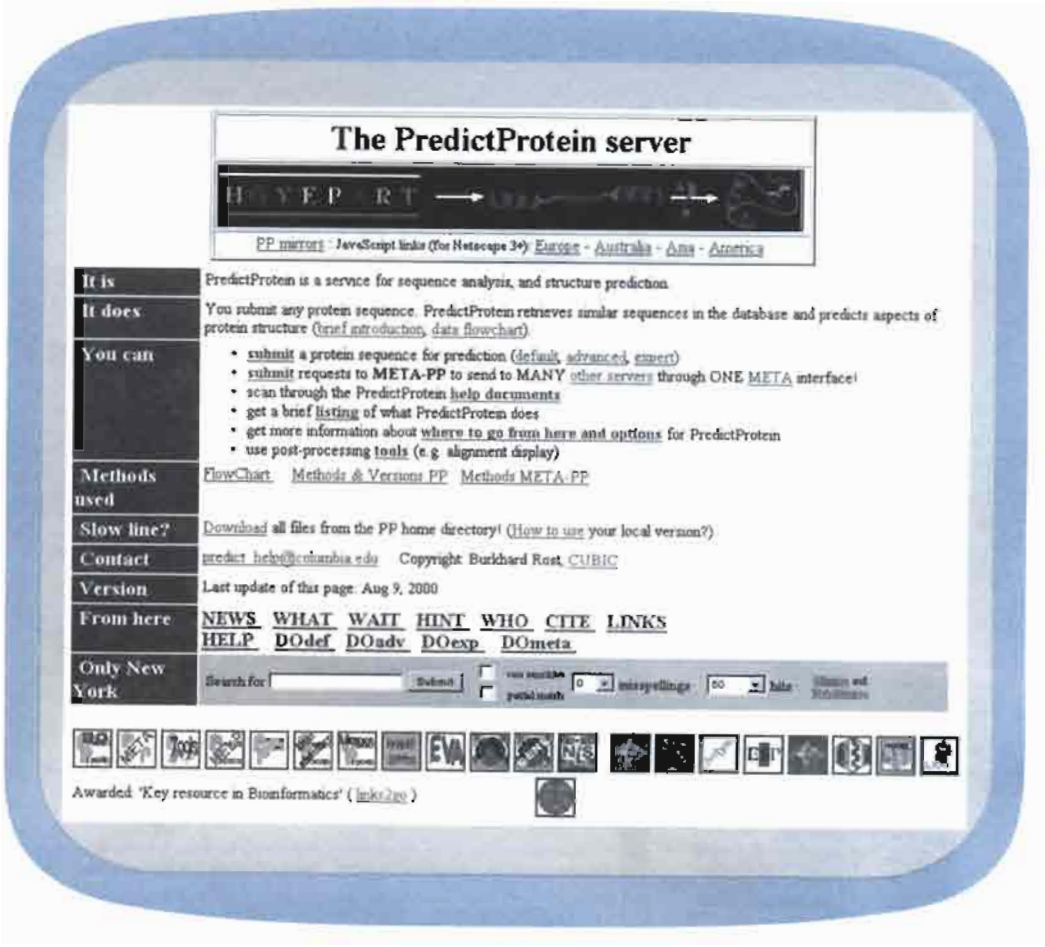


FIGURE 9.38. *The PredictProtein server offers protein prediction including comparative protein modeling. The main URL is* ▸*http://cubic.bioc.columbia.edu/predictprotein/, and there are about 20 mirror sites worldwide. PredictProtein offers database searches and predictions of protein secondary structure, residue solvent accessibility, transmembrane helices, protein globularity, and other features.*

Wolf et al. (1999) combined PSI-BLAST searches of proteins from 13 completed genomes (including bacteria, archaea, and eukaryotes) with structural data from SCOP and predicted folds for 20–30% of all proteins.

The "Rosetta Stone" method is one of the most successful ab initio strategies (Simons et al., 2001). The target protein is evaluated in fragments of nine amino acids. These fragments are compared to known structures in PDB. From this analysis, structures can be inferred for the entire peptide chain. Bonneau et al. (2002) used the Rosetta Stone method to model the structure of all Pfam-A sequence families (Chapter 10) for which three-dimensional structures are unknown. By calibrating their method on known structures, they estimated that for 60% of the proteins studied (80 of 131), one of the top five ranked models successfully predicted the structure within 6.0 Å RMSD.

PROTEIN STRUCTURE PREDICTION AND LIMITS OF PROTEIN FOLD SPACE

What are the prospects for identifying all folds that occur in nature? There may be a total of between 1000 and 5000 folds in all life forms (Brenner et al., 1997; Burley, 2000). PDB catalogs about 700 distinct protein folds, some of which are represented by many protein sequence families. For example, the TIM barrel (Fig. 9.23) occurs in over 20 distinct families. The average protein has about two globular domains,

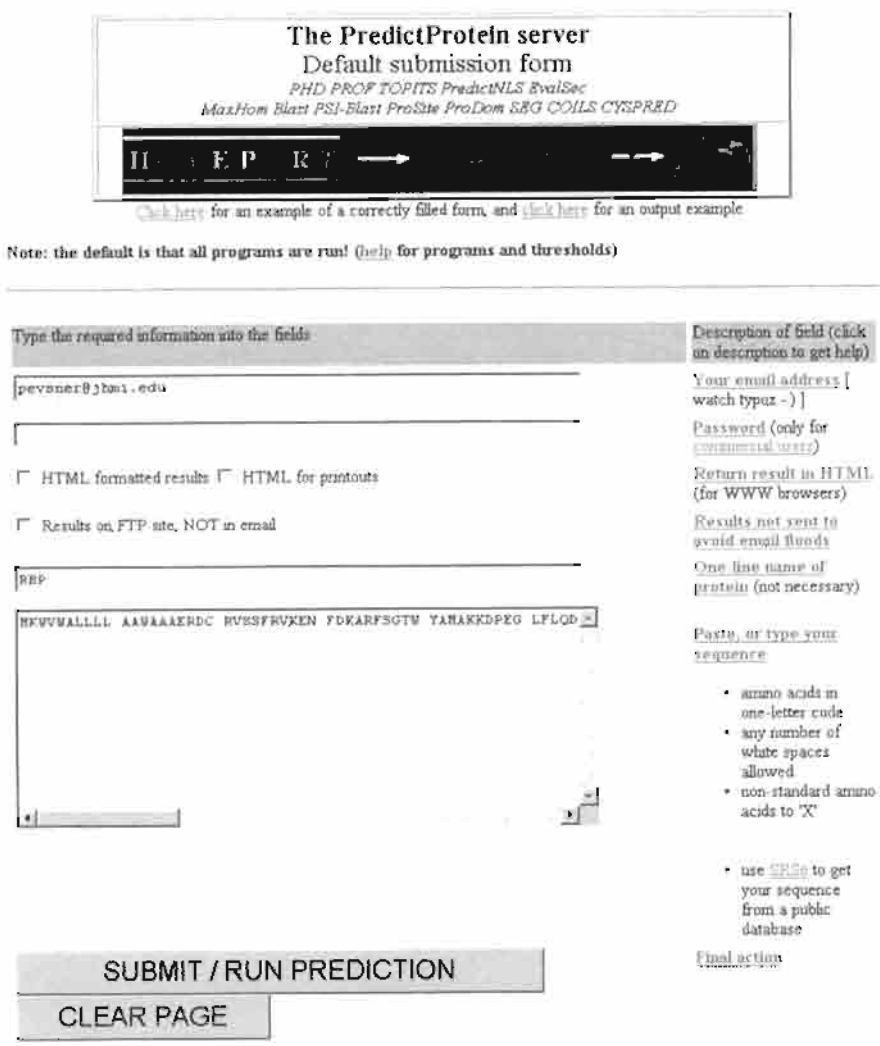


FIGURE 9.39. *The PredictProtein server includes a simple query form for protein submissions.*

and the average domain is 153 ± 87 residues (Orengo et al., 1999). As more protein structures are solved, the proportion of novel folds is declining. Thus there is a prospect that we are gradually approaching the goal of full coverage of protein sequence space (Burley, 2000).

The state of the art of protein prediction is assessed by the structural genomics community at Critical Assessment of Techniques for Protein Structure Prediction (CASP) (Zemla et al., 2001; Venclovas et al., 2001). This is a double-blind structure prediction experiment (or competition). Dozens of teams from around the world participate. Each team's goal is to correctly predict the structures of a group of dozens of target protein sequences. The structure for each of these targets is known but unpublished. The CASP organizers assess the predictions (without knowledge of the teams' identities). At CASP4 (December 2000), the largest source of error in the process of structure prediction was in correct sequence alignment.

Information on the CASP experiments is available at ▸http://predictioncenter.llnl.gov/. In CASP4, the targets included pig β-lactoglobulin (target T123).

Protein Structure and Disease

The linear sequence of amino acids specifies the three-dimensional structure of a protein. A change in even a single amino acid can cause a profound disruption in structure. For example, cystic fibrosis is caused by mutations in the gene encoding

TABLE 9-5 Examples of Diseases for Which Subtle Change in Protein Sequence Leads to Change in Structure

Disease	Gene/Protein	LocusID	RefSeq
Cystic fibrosis	CFTR	1080	NP_000483
Sickle cell anemia	Hemoglobin beta	3043	NP_000509
"Mad cow" disease (BSE)	Prion protein	5621	NP_000302
Alzheimer disease	Amyloid precursor protein	351	NP_000475

Abbreviations: CFTR, cystic fibrosis transmembrane regulator; BSE, bovine spongiform encephalopathy.

cystic fibrosis transmembrane regulator (CFTR) (Ratjen and Döring, 2003). The most common mutation is *ΔF508*, a deletion of a phenylalanine at position 508. The consequence of removing this residue is to alter the alpha helical content of the protein. This in some way impairs the ability of the CFTR protein to traffic through the secretory pathway to its normal location on the plasma membrane of lung epithelial cells.

Changes in protein sequence that are associated with disease do not necessarily cause large changes in protein structure. An example is provided by sickle cell anemia, the most common inherited blood disorder. It is caused by mutations in the gene encoding hemoglobin beta (also called beta globin) on chromosome 11p15.4. Adult hemoglobin is a tetramer consisting of two alpha chains and two beta chains. The protein carries oxygen in blood from the lungs to various parts of the body. A substitution of a valine for a normally occurring glutamic acid residue forms a hydrophobic patch on the surface of the beta globin, leading to clumping of many hemoglobin molecules.

You can access a brief definition of the hemoglobin chains at LocusLink (LocusID: 3043). You can also find a link there to Online Mendelian Inheritance in Man, which provides a detailed description of the clinical and molecular consequences of globin gene mutations. We will discuss sickle cell anemia in Chapter 18.

Several examples of proteins associated with human disease are presented in Table 9.5, including CFTR and beta globin. We will examine their normal and mutant structures in the problems at the end of this chapter.

PERSPECTIVE

The aim of structural genomics is to define structures that span the entire space of protein folds. This project has many parallels to the Human Genome Project. Both are ambitious endeavors that require the international cooperation of many laboratories. Both involve central repositories for the deposit of raw data, and in each the growth of the databases is exponential.

It is realistic to expect that the great majority of protein folds will be defined in the near future. Each year, the proportion of novel folds declines rapidly. A number of lessons are emerging:

- Proteins assume a limited number of folds.
- A single three-dimensional fold may be used by proteins to perform entirely distinct functions.
- The same function may be performed by proteins using entirely different folds.

PITFALLS

One of the great mysteries of biology is how the linear amino acid sequence of a protein folds quickly into the correct three-dimensional conformation. For the computational biologist, it is extraordinarily difficult to predict a three-dimensional

structure from primary data alone. Thus ab initio folding approaches are less successful than comparative modeling techniques.

WEB RESOURCES

TABLE 9-6 Partial List of Protein Structure Databases Linked to Protein Data Bank

Database	Comment	URL
3dee	Structural domain definitions	▸http://jura.ebi.ac.uk:8080/3Dee/help/help_intro.html
BMCD	Crystallization information about biomacromoleucles	▸http://wwwbmcd.nist.gov:8080/bmcd/bmcd.html
CATH	Protein fold classifiction	▸http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html
CE	Complete PDB and representative structure comparison and alignments	▸http://cl.sdsc.edu/ce.html
DSSP	Secondary-structure classification	▸http://www.cmbi.kun.nl/swift/dssp/
Enzyme Structures Databases	Enzyme classifications and nomenclature	▸http://www.biochem.ucl.ac.uk/bsm/
FSSP	Structurally similar families	▸http://www.ebi.ac.uk/dali/fssp
GRASS	Graphical representation and analysis	▸http://honiglab.cpmc.columbia.edu/cgi-bin/GRASS/surfserv_enter.cgi
HSSP	Homology-derived secondary structures	▸http://www.cmbi.kun.nl/swift/hssp
MMDB	Database of three-dimensional structures	▸http://www.nebi.nlm.nih.gov/structure/
MEDLINE	Direct access to MEDLINE at NCBI	▸http://www.nlm.nih.gov/databases/medline.html
NDB	Database of three-dimensional nucleic acid structures	▸http://ndbserver.rutgers.edu/
PDBSum	Summary information about protein structures	▸http://www.biochem.ucl.ac.uk/bsm/pdbsum/
SCOP	Structure classifications	▸http://scop.mrc-lmb.cam.ac.uk/scop/
Tops	Protein strucutre motif comparisons topological diagram	▸http://tops.ebi.ac.uk:80/tops/
VAST	Vector Alignment Search Tool (NCBI)	▸http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml
Whatcheck	Protein structure checks	▸http://biotech.ebi.ac.uk:8400/

Source: Modified from Berman et al. (2000) and PDB (▸http://resb.org/pdb/links.html)

DISCUSSION QUESTIONS

[9-1] The Protein Data Bank (PDB) is the central repository of protein structure data. What do databases such as SCOP and CATH offer that PDB lacks?

[9-2] A general rule is that protein structure evolves more slowly than primary amino sequence. Thus, two proteins can have only limited amino acid sequence identity while sharing highly similar structures. (A good example of this is the lipocalins, where retinol-binding protein, odorant-binding protein, and β-lactoglobulin share highly related structures with low sequence identity.) Are there likely to be exceptions to this general rule?

PROBLEMS

[9-1] View the structure of a protein:

- Go to NCBI Entrez Structures and select a lipocalin. You can access this from the main NCBI page by going to "structure." Alternatively, in Entrez you can type a query, select "limits," and restrict the output to PDB. If you select "odorant-binding protein," there are entries for

odorant-binding proteins from several different species. From cow, there are entries deposited independently from different research groups (e.g., PDB identifiers 1OBP, 1PBO).

- Select "View 3D Structure" in the MMDB web page. Explore the links on the page. Click "View/Save Structure." (You may need to click to download the Cn3D software.)
- Two windows open: the Cn3D viewer and the 1D-viewer. Click on each of these, and notice how they are interconnected. Change the "style" of the Cn3D viewer. Identify the α helices and β sheets of the protein.

[9-2] Compare the structures of two lipocalins:

- Go back to the MMDB page for 1OBP and select "Structure neighbors." (This can be accessed by mousing over the protein graphic.) You are now looking at the NCBI VAST (Vector Alignment Search Tool) site. There is a list of proteins related to OBP. Select one or two other proteins, such as β-lactoglobulin or retinol-binding protein, by clicking on the box(es) to the left. Now view/save the alignments.
- Notice that two windows open up: Cn3D and DDV (the two-dimensional viewer). Again explore the relationship between these two visualization tools. What are the similarities between the proteins you are comparing? What are their differences? Highlight the regions of conserved amino acids both in the alignment viewer and the graphical viewer. Where are the invariant GXW residues located?

[9-3] Mutations in the beta chain of hemoglobin (gene symbol HBB; also called beta globin) can cause sickle cell anemia or other diseases. Try to find the PDB accession numbers for both normal hemoglobin and a mutated form. Try the following:

- The NCBI Structure page
- The PDB
- CATH or SCOP
- A blastp search against the PDB at the NCBI website

[9-4] Sickle-cell anemia is caused by a specific mutation in HBB, *E6V* (i.e., a glutamic acid residue at amino acid position 6 is substituted with a valine). As a consequence of this mutation, hemoglobin tetramers can clump together. This causes the entire red blood cell to deform, adopting a sickled shape. Use PDB identifiers 1HBB or 4HHB for wild-type hemoglobin and 2HBS for a mutant form. Compare the structures using the VAST tool at NCBI. Is the glutamate at position 6 on the surface of the protein or is it buried inside? Does the mutation to a valine cause a change in the predicted secondary or tertiary structure of the protein?

[9-5] The most common cause of cystic fibrosis is $\Delta F508$. This is a deletion in the gene encoding cystic fibrosis transmembrane regulator (CFTR) in which a phenylalanine at position 508 is removed. Massiah et al. (1999) used NMR to solve the structure of a 26-amino-acid peptide corresponding to a portion of the wild-type CFTR (PDB accession 1CKY) and a 25-amino-acid peptide lacking the phenylalanine residue (1CKZ). These two peptides are too small to compare in the NCBI VAST program. Use other programs to predict whether the peptides differ in their secondary-structure content.

SELF-TEST QUIZ

[9-1] In comparing two homologous but distantly related proteins:

(a) They tend to share more three-dimensional structure features in common than percent amino acid identity.
(b) They tend to share more percent amino acid identity in common than three-dimensional structure features.
(c) They tend to share three-dimensional structure features and percent amino acid identity to a comparable extent.
(d) It is not reasonable to generalize about the extent to which they share three-dimensional structure features and percent amino acid identity.

[9-2] Protein secondary structure prediction algorithms typically calculate the likelihood that a protein:

(a) Forms α helices
(b) Forms α helices and β sheets
(c) Forms α helices, β sheets, and coils
(d) Forms α helices, β sheets, coils, and multimers

[9-3] An advantage of X-ray crystallography relative to NMR for structure determination is that using X-ray crystallography:

(a) It is easier to solve the structure of transmembrane domain-containing proteins.
(b) It is easier to grow crystals than to prepare samples for NMR.
(c) It is easier to interpret diffraction data.
(d) It is easier to determine the structures of large proteins.

[9-4] The Protein Data Bank (PDB):

(a) Functions primarily as the major worldwide repository of macromolecular secondary structures.
(b) Contains approximately as many structures as there are protein sequences in SwissProt/TrEMBL.
(c) Includes data on proteins, DNA–protein complexes as well as carbohydrates.
(d) Is operated jointly by the NCBI and EBI.

[9-5] The NCBI VAST algorithm:

(a) Is a web browser tool for the visualization of related protein structures by threading.

(b) Is a visualization tool that allows the simultaneous comparison of as many as two structures.
(c) Allows searches of all the NCBI structure database with queries that have known structures (i.e., having PDB accession numbers), but this tool is not useful for the analysis of uncharacterized structures.
(d) Allows searches of all the NCBI structure database entries against each other and provides a list of "structure neighbors" for a given query.

[9-6] Cn3D is a molecular structure viewer at NCBI. It features
(a) A menu-driven program linked to automated homology modeling
(b) A command line interface useful for a variety of structure analyses
(c) A structure viewer that is accompanied by a sequence viewer
(d) A structure viewer that allows stereoscopic viewing of structure images

[9-7] The CATH database offers a hierarchical classification of protein structures. The first three levels, class (C), architecture (A), and topology (T), all describe:
(a) Protein tertiary structure (e.g., tertiary structure composition, packing, shape, orientation, and connectivity)
(b) Protein secondary structure (e.g., secondary structure composition, packing, shape, orientation, and connectivity)
(c) Protein domain structure
(d) Protein superfamilies grouped according to homologous domains

[9-8] Homology modeling may be distinguished from ab initio prediction because:
(a) Homology modeling requires a model to be built.
(b) Homology modeling requires alignment of a target to a template.
(c) Homology modeling is usefully applied to any protein sequence.
(d) The accuracy of homology modeling is independent of the percent identity between the target and the template.

[9-9] You have a protein sequence, and you want to quickly predict its structure. After performing BLAST and PSI-BLAST searches, you identify the most closely related proteins with a known structure as several having 17% amino acid identity to your protein. Which of these options is best?
(a) X-ray crystallography
(b) NMR
(c) Submitting your sequence to a protein structure prediction server that performs homology modeling
(d) Submitting your sequence to a protein structure prediction server that performs ab initio modeling

Suggested Reading

There are many superb overviews of structural genomics and protein structure prediction. These include articles by Janet Thornton, Christine Orengo, and colleagues (Thornton et al., 2000; Todd et al., 2001; Orengo et al., 2001); Stephen Burley (Burley et al., 1999; Burley, 2000); David T. Jones of Brunel University (Jones, 2001); and many others (Teichmann et al., 1999; Holm, 1998; Holm and Sander, 1996b; Domingues et al., 2000). Most of these articles discuss the relation between protein structure and function, and evaluate methods for protein structure prediction. Another recommended overview, by Mark Gerstein (2000), is part of a Nature Structural Biology supplement on structural genomics (▸http://structbio.nature.com). For comparative protein structure modeling, see the review by Andrej Sali and colleagues (Marti-Renom et al., 2000).

A variety of papers offer overviews of the specific databases described in this chapter. For the Protein Data Bank see articles by Berman and colleagues (Berman et al., 2002; Westbrook et al., 2002). For the CATH database, see the overview from Christine Orengo and colleagues (Pearl et al., 2000). SCOP is described by Murzin et al. (1995), and for Dali and FSSP, see Dietmann et al. (2001) and Holm and Sander (1996a).

References

Baker, D., and Sali, A. Protein structure prediction and structural genomics. *Science* **294,** 93–96 (2001).

Berman, H. M., et al. The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

Berman, H. M., et al. The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.* **58,** 899–907 (2002).

Bonneau R., Strauss C. E., Rohl C. A., Chivian D., Bradley P., Malmstrom L., Robertson T., Baker D. De novo prediction of three-dimensional structures for major protein families. *J. Mol. Biol.* **322,** 65–78 (2002).

Brenner, S. E., Chothia, C., and Hubbard, T. J. Population statistics of protein structures: Lessons from structural classifications. *Curr. Opin. Struct. Biol.* **7,** 369–376 (1997).

Brenner, S. E. Target selection for structural genomics. *Nat. Struct. Biol.* **7** (Suppl.), 967–969 (2000).

Burley, S. K., et al. Structural genomics: Beyond the human genome project. *Nat. Genet.* **23,** 151–157 (1999).

Burley, S. K. An overview of structural genomics. *Nat. Struct. Biol.* **7** (Suppl.), 932–934 (2000).

Bystroff, C., and Shao, Y. Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics.* **18** Suppl 1, S54–61 (2002).

Chang, G., and Roth, C. B. Structure of MsbA from E. coli: A Homolog of the Multidrug Resistance ATP Binding Cassette (ABC) Transporters. *Science* **293,** 1793–800 (2001).

Chou, P. Y., and Fasman, G. D. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* **47,** 45–148 (1978).

Dietmann, S., et al. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.* **29,** 55–57 (2001).

Domingues, F. S., Koppensteiner, W. A., and Sippl, M. J. The role of protein structure in genomics. *FEBS Lett.* **476,** 98–102 (2000).

Flower, D. R., North, A. C., and Sansom, C. E. The lipocalin protein family: Structural and sequence overview. *Biochim. Biophys. Acta* **1482,** 9–24 (2000).

Garnier, J., Gibrat, J. F., and Robson, B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **266,** 540–553 (1996).

Garnier, J., Osguthorpe, D. J., and Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120,** 97–120 (1978).

Gerstein, M. Integrative database analysis in structural genomics. *Nat. Struct. Biol.* **7** (Suppl.), 960–963 (2000).

Gerstein, M., and Levitt, M. A structural census of the current population of protein sequences. *Proc. Natl. Acad. Sci. USA* **94,** 11911–11916 (1997).

Holm, L., and Sander, C. The FSSP database: Fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.* **24,** 206–209 (1996a).

Holm, L., and Sander, C. Mapping the protein universe. *Science* **273,** 595–603 (1996b).

Holm, L. Unification of protein families. *Curr. Opin. Struct. Biol.* **8,** 372–379 (1998).

Jones, D. T. Protein structure prediction in genomics. *Brief. Bioinform.* **2,** 111–125 (2001).

Koonin, E. V., Wolf, Y. I., and Karev, G. P. The structure of the protein universe and genome evolution. *Nature* **420,** 218–223 (2002).

Marti-Renom, M. A., et al. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29,** 291–325 (2000).

Massiah, M. A., Ko, Y. H., Pedersen, P. L., and Mildvan, A. S. Cystic fibrosis transmembrane conductance regulator: Solution structures of peptides based on the Phe508 region, the most common site of disease-causing DeltaF508 mutation. *Biochemistry* **38,** 7453–7461 (1999).

McGuffin, L. J., and Jones, D. T. Targeting novel folds for structural genomics. *Proteins* **48,** 44–52 (2002).

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247,** 536–540 (1995).

Orengo, C. A., et al. CATH—a hierarchic classification of protein domain structures. *Structure* **5,** 1093–1108 (1997).

Orengo, C. A., Bray, J. E., Hubbard, T., LoConte, L., and Sillitoe, I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. *Proteins* (Suppl.) **3,** 149–170 (1999).

Orengo, C. A., Sillitoe, I., Reeves, G., and Pearl, F. M. Review: What can structural classifications reveal about protein evolution? *J. Struct. Biol.* **134,** 145–165 (2001).

Osguthorpe, D. J. Ab initio protein folding. *Curr. Opin. Struct. Biol.* **10,** 146–152 (2000).

Pearl, F. M., et al. Assigning genomic sequences to CATH. *Nucleic Acids Res.* **28,** 277–282 (2000).

Pearson, W. R., and Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448 (1988).

Przybylski, D., and Rost, B. Alignments grow, secondary structure prediction improves. *Proteins* **46,** 197–205 (2002).

Ratjen, F., and Döring, G. Cystic fibrosis. *Lancet* **361,** 681–689 (2003).

Rost, B., and Sander, C. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232,** 584–99 (1993a).

Rost, B., and Sander, C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci. USA* **90,** 7558–7562 (1993b).

Rost, B. Review: Protein secondary structure prediction continues to rise. *J. Struct. Biol.* **134,** 204–218 (2001).

Shortle, D. Prediction of protein structure. *Curr. Biol.* **10,** R49–51 (2000).

Simons, K. T., Strauss, C., and Baker, D. Prospects for ab initio protein structural genomics. *J. Mol. Biol.* **306,** 1191–1199 (2001).

Taylor, W. R., and Orengo, C. A. Protein structure alignment. *J. Mol. Biol.* **208,** 1–22 (1989a).

Taylor, W. R., and Orengo, C. A. A holistic approach to protein structure alignment. *Protein Eng.* **2,** 505–519 (1989b).

Teichmann, S. A., Chothia, C., and Gerstein, M. Advances in structural genomics. *Curr. Opin. Struct. Biol.* **9,** 390–399 (1999).

Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N., and Orengo, C. A. From structure to function: Approaches and limitations. *Nat. Struct. Biol.* **7** (Suppl.), 991–994 (2000).

Todd, A. E., Orengo, C. A., and Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307,** 1113–1143 (2001).

Venclovas, C., Zemla, A., Fidelis, K., and Moult, J. Comparison of

performance in successive CASP experiments. *Proteins* (Suppl.), **5,** 163–170 (2001).

Wang, Y., et al. MMDB: 3D structure data in Entrez. *Nucleic Acids Res.* **28,** 243–245 (2000).

Westbrook, J., et al. The Protein Data Bank: Unifying the archive. *Nucleic Acids Res.* **30,** 245–248 (2002).

Wolf, Y. I., Brenner, S. E., Bash, P. A., and Koonin, E. V. Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **9,** 17–26 (1999).

Wood, T. C., and Pearson, W. R. Evolution of protein sequences and structures. *J. Mol. Biol.* **291,** 977–95 (1999).

Zemla, A., Venclovas, Moult, J., and Fidelis, K. Processing and evaluation of predictions in CASP4. *Proteins* (Suppl.), 13–21 (2001).

Tabelle 2

VERGLEICH EINER HOMOLOGEN REGION IN CYTOCHROM-C VERSCHIEDENER HERKUNFT

Rind	...-Val-Glu(NH_2)-Lys-CyS-Ala-Glu(NH_2)-CyS-His-Thr-Val-Glu-Lys-...
Pferd	.....-Lys-CyS-Ala-Glu(NH_2)-CyS-His-Thr-Val-Glu-Lys-...
Schwein	.....-Lys-CyS-Ala-Glu(NH_2)-CyS-His-Thr-Val-Glu-Lys-...
Lachs	...-Val-Glu(NH_2)-Lys-CyS-Ala-Glu(NH_2)-CyS-His-Thr-Val-Glu- .....
Huhn	...-Val-Glu(NH_2)-Lys-CyS-*Ser*-Glu(NH_2)-CyS-His-Thr-Val-Glu-.....
Seiden-spinner	...-Val-Glu(NH_2)-*Arg*-CyS-Ala-Glu(NH_2)-Cys-His-Thr-Val-Glu-.....
Hefe	...-*Phe-Lys-Thr*-------- *Arg*--CyS-*Glu*--*Leu*--------CyS-His-Thr-Val-Glu-.....

als mit der in Säugetier-Cytochromen ermittelten identisch, in Hühner-Cytochrom-c hingegen war an die Stelle eines Alanin-Restes ein Serin-Rest getreten. Ein Invertebraten-Cytochrom-c, das des Seidenspinners Bombyx

(a)

Tabelle 3

ARTUNTERSCHIEDE IN INSULIN (NACH HARRIS, SANGER UND NAUGHTON, 1956)

Rind	...-CyS-*Ala-Ser-Val*-CyS-...
Schwein	...-CyS-*Thr-Ser-Ileu*-CyS-...
Schaf	...-CyS-*Ala-Gly-Val*-CyS-...
Pferd	...-CyS-*Thr-Gly-Ileu*-CyS-...
Wal	...-CyS-*Thr-Ser-Ileu*-CyS-...

(b)

As the linear amino acid sequences of proteins were determined in the 1950s and 1960s, it became of obvious interest to try to align them. (a) *Hans Tuppy (1958, p. 71) described the alignment of cytochromes c from Rind (beef), Pferd (horse), Schwein (pig), Lachs (salmon), Huhn (chicken), Seiden-spinner (silkworm), and Hefe (yeast). This alignment showed that even though gaps had to be introduced, protein sequences from organisms as distantly related as mammals and yeast could still be aligned.* (b) *Tuppy (1958, p. 73) also described an alignment of insulin amino acid sequences from beef, pig, Schaf (sheep), horse, and Wal (whale). In this case, he noted the lack of conservation of several amino acid residues in a region between two cysteine residues. For more details on the alignment of insulins, see Figure 11.1. Used with permission.*

10

Multiple Sequence Alignment

INTRODUCTION

When we consider a protein (or gene), one of the most fundamental questions is what other proteins are related. Biological sequences often occur in families. These families may consist of related genes within an organism (paralogs), sequences within a population (e.g., polymorphic variants), or genes in other species (orthologs). Sequences diverge from each other for reasons such as duplication within a genome or speciation leading to the existence of orthologs. We have studied pairwise comparisons of two protein (or DNA) sequences (Chapter 3), and we have also seen multiple related sequences as the output of a BLAST or other database search (Chapters 4 and 5). We also explored protein domains (Chapter 8), and we saw that protein structure databases often display homologous protein sequences in the form of a multiple sequence alignment (Chapter 9). In this chapter, we will consider the general problem of multiple sequence alignment. While multiple sequence alignment is commonly performed for both protein and DNA sequences, most databases consist of protein families only.

Multiple sequence alignments are of great interest because homologous sequences often retain similar structures and functions. By introducing sequences into a multiple alignment, we can define members of a gene or protein family. If we know a feature of one of the proteins (e.g., RBP4 transports a hydrophobic ligand), then when we identify homologous proteins, we can predict that they may have similar function. The overwhelming majority of proteins have been identified through the sequencing of genomic DNA or cDNA. Thus the function of most proteins is

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

assigned on the basis of homology to other known proteins rather than on the basis of results from biochemical or cell biological (functional) assays.

Definition of Multiple Sequence Alignment

Domains or motifs that characterize a protein family are defined by the existence of a multiple sequence alignment of a group of homologous sequences. A multiple sequence alignment is a collection of three or more protein (or nucleic acid) sequences that are partially or completely aligned. Homologous residues are aligned in columns across the length of the sequences. These aligned residues are homologous in an evolutionary sense: They are presumably derived from a common ancestor. The residues in each column are also presumed to be homologous in a structural sense: Aligned residues tend to occupy corresponding positions in the three-dimensional structure of each aligned protein.

Multiple sequence alignments are easy to generate, even by eye, for a group of very closely related protein (or DNA) sequences. We have seen examples of alignments of closely related sequences in Figures 3.7 and 3.8 (GAPDH and RBPs). As soon as the sequences exhibit some divergence, the problem of multiple alignment becomes extraordinarily difficult to solve. In particular, the number and location of gaps is difficult to assess. We saw an example of this with human lipocalins (Fig. 3.9). In this chapter we will compare alignment results using five distantly related lipocalins as well as five closely related RBP orthologs.

There is not necessarily one "correct" alignment of a protein family. This is because while protein structures tend to evolve over time, protein sequences generally evolve even more rapidly than structures. Looking at the sequences of RBP4 and β-lactoglobulin, we saw that they share only 26% amino acid identity (Fig. 3.5), but the three-dimensional structures are nearly identical (Figs. 3.1 and 9.18). In creating a multiple sequence alignment, it may be impossible to identify the amino acid residues that should be aligned with each other as defined by the three-dimensional structures of the proteins in the family. We typically do not have high-resolution structural data available, and we rely on sequence data to generate the alignment. Similarly, we often do not have functional data to identify domains (such as the specific amino acids that form the catalytic site of an enzyme), so again we rely on sequence data. It is possible to compare the results of multiple sequence alignments that are generated solely from sequence data and to then examine known structures for those proteins. For a given pair of divergent but significantly related protein sequences (e.g., for two proteins sharing 30% amino acid identity), Chothia and Lesk (1986) found that about 50% of the individual amino acid residues are superposable in the two structures.

Aligned columns of amino acid residues characterize a multiple sequence alignment. This alignment may be determined because of features of the amino acids such as the following:

- There are highly conserved residues such as cysteines that are involved in forming disulfide bridges.
- There are conserved motifs such as a transmembrane domain or an immunoglobulin domain. We saw examples of protein domains and motifs (such as the PROSITE dictionary) in Chapter 8.
- There are conserved features of the secondary structure of the proteins, such as residues that contribute to α helices, β sheets, or transitional domains.
- There are regions that show consistent patterns of insertions or deletions.

Typical Uses and Practical Strategies of Multiple Sequence Alignment

When and why are multiple sequence alignments used?

- If a protein (or gene) you are studying is related to a larger group of proteins, this group membership can often provide insight into the likely function, structure, and evolution of that protein.
- Most protein families have distantly related members. Multiple sequence alignment is a far more sensitive method than pairwise alignment to detect homologs (Park et al., 1998).
- When one examines the output of any database search (such as a BLAST search), a multiple sequence alignment format can be extremely useful to reveal conserved residues or motifs in the output.
- If one is studying cDNA clones, it is common practice to sequence them. Multiple sequence alignment can show whether there are any discrepancies in the sequences.
- Analysis of population data can provide insight into many biological questions involving evolution, structure, and function. The PopSet portion of Entrez (described below) contains nucleotide (and protein) population data sets that are viewed as multiple alignments.
- When the complete genome of any organism is sequenced, a major portion of the analysis consists of defining the protein families to which all the gene products belong. Database searches effectively perform multiple sequence alignments, comparing each novel protein (or gene) to the families of all other known genes.
- We will see in Chapter 11 how phylogeny algorithms begin with multiple sequence alignments as the raw data with which to generate trees. The most critical part of making a tree is to produce an optimal alignment.
- The regulatory regions of many genes contain consensus sequences for transcription factor-binding sites.

Feng and Doolittle's Progressive Sequence Alignment

A variety of approaches have been developed to perform multiple sequence alignment. The most commonly used algorithms that produce multiple alignments are derived from the progressive alignment method of Da-Fei Feng and Russell Doolittle (1987, 1990). It is called "progressive"because the strategy entails calculating pairwise sequence alignment scores between all the proteins (or DNA sequences) being aligned, then beginning the alignment with the two closest sequences and progressively adding more sequences to the alignment.

We can illustrate the procedure by aligning five lipocalins, selected from Entrez and pasted into a text document in the FASTA format. These five protein sequences are entered into the ClustalW program (Fig. 10.1). This is one of the most popular programs for performing multiple sequence alignments (see below) (Thompson et al., 1994).

ClustalW is accessed online at ▸http://www2.ebi.ac.uk/clustalw/, where it is hosted by the European Bioinformatics Institute.

Feng and Doolittle's progressive alignment procedure occurs in three stages. These are outlined for the case of five distantly related lipocalins (Figs. 10.2 and 10.3) as well as five closely related RBPs (Figs. 10.4 and 10.5):

1. In stage 1, the global alignment approach of Needleman and Wunsch (1970; Chapter 3) is used to create pairwise alignments of every protein that is to be

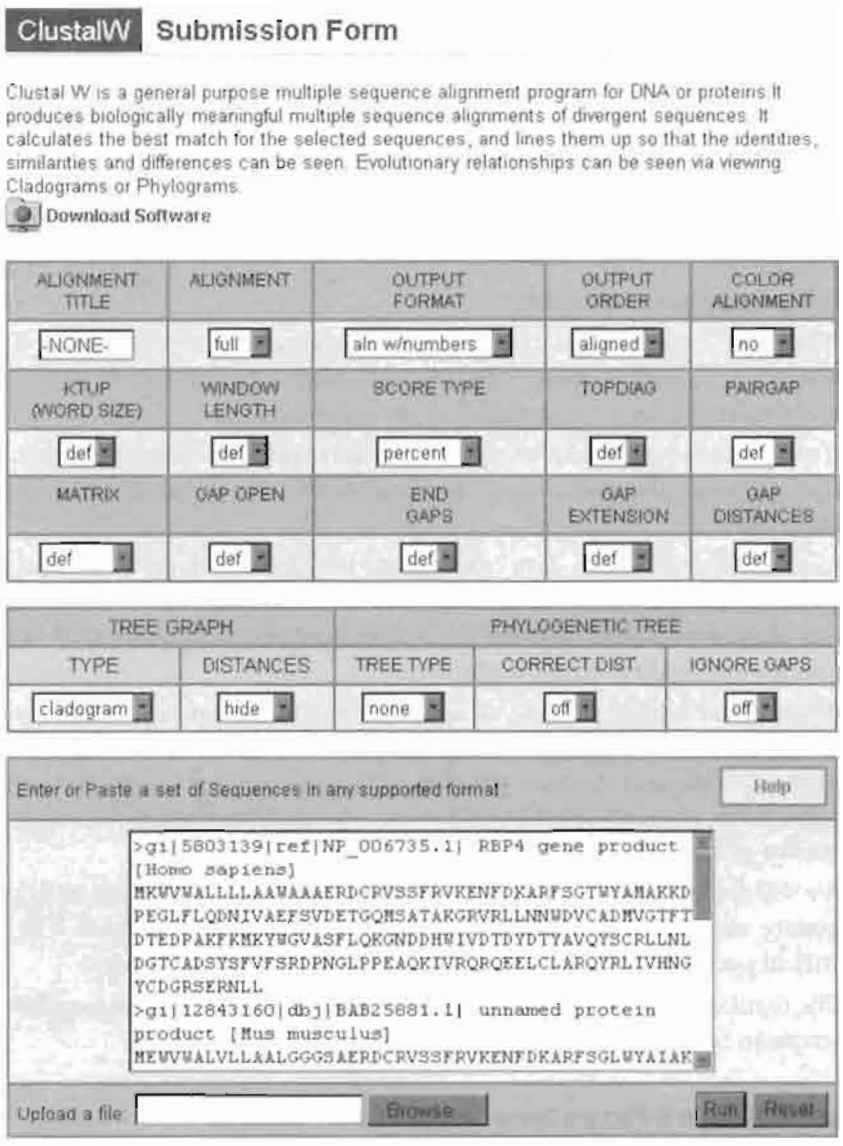


FIGURE 10.1. *Multiple sequence alignment of five lipocalins using the ClustalW server at EBI (http://www2.ebi.ac.uk/clustalw/). The proteins were retinol binding protein 4 from human (NP_006735) and mouse (BAB25881), human apolipoprotein D (NP_001638), mouse major urinary protein 4 (P11590), and* E. coli *outer membrane lipocalin BLC (P39281). Each protein was pasted in using the FASTA format from Entrez (NCBI).*

included in a multiple sequence alignment (Fig. 10.2, stage 1). As shown in the figure, for an alignment of five sequences, 10 pairwise alignment scores are generated.

For N sequences that are multiply aligned, the number of pairwise alignments that must be calculated for the initial matrix equals $\frac{1}{2}(N-1)(N)$. For five proteins, 10 pairwise alignments are made.

Algorithms that perform pairwise alignments generate raw similarity scores. Many progressive sequence alignment algorithms including ClustalW use a distance matrix rather than a similarity matrix to describe the relatedness of the proteins. The conversion of similarity scores for each pair of sequences to distance scores is outlined in Box 10.1. The purpose of generating distance measures is to generate a guide tree (stage 2, below) to construct the alignment.

In our example, note that the best pairwise alignment score is for human versus mouse RBP4 (Fig. 10.2, arrow 1). As we will describe below, most multiple sequence alignments employ global pairwise alignments, although some use local alignment. For a group of closely related RBPs, all have similarly high scores (Fig. 10.4).

Stage 1: generate a series of pairwise sequence alignments

```
Sequence format is Pearson
Sequence 1: gi|5803139|ref|NP_006735.1|          199 aa
Sequence 2: gi|12843160|dbj|BAB25881.1|          201 aa
Sequence 3: gi|4502163|ref|NP_001638.1|          189 aa
Sequence 4: gi|127528|sp|P11590|MUP4_MOUSE       178 aa
Sequence 5: gi|732003|sp|P39281|BLC_ECOLI        177 aa

Start of Pairwise alignments
Aligning...
Sequences (1:2) Aligned. Score:  84 <-- 1
Sequences (1:3) Aligned. Score:  14
Sequences (1:4) Aligned. Score:  8
Sequences (1:5) Aligned. Score:  12
Sequences (2:3) Aligned. Score:  17
Sequences (2:4) Aligned. Score:  9
Sequences (2:5) Aligned. Score:  19
Sequences (3:4) Aligned. Score:  9
Sequences (3:5) Aligned. Score:  27
Sequences (4:5) Aligned. Score:  7
```

Stage 2: create a guide tree, calculated from the distance matrix

```
(
(
(
gi|5803139|ref|NP_006735.1|:0.09385,
gi|12843160|dbj|BAB25881.1|:0.05691)
:0.34199,
gi|127528|sp|P11590|MUP4_MOUSE:0.48994)
:0.06118,
gi|4502163|ref|NP_001638.1|:0.35805,
gi|732003|sp|P39281|BLC_ECOLI:0.36511);
```

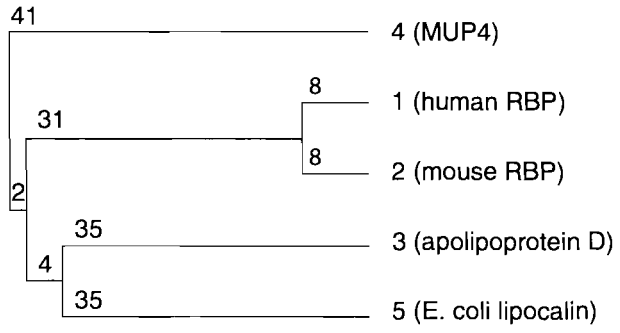


FIGURE 10.2. *Progressive alignment method of Feng and Doolittle (1987) used by many multiple alignment programs such as ClustalW. In stage 1, a series of pairwise alignments is generated for five distantly related lipocalins (see Fig. 10.1). Note that the best score is for an alignment of human versus mouse RBP (score = 84; arrow 1). In stage 2, a guide tree is calculated describing the relationships of the five sequences based upon their pairwise alignment scores. A graphical representation of the guide tree is shown using the JalView tool at the ClustalW web server. Branch lengths (rounded off) reflect distances between sequences and are indicated on the tree; compare to Figure 10.4.*

2. In the second stage, a guide tree is calculated from the similarity (or distance) matrix. There are two principal ways to construct a guide tree: the unweighted pair group method of arithmetic averages (UPGMA) and the neighbor-joining method. We will define these algorithms in Chapter 11. The two main features of a tree are its topology (branching order) and branch lengths (which can be drawn so that they are proportional to evolutionary distance). Thus the tree reflects the relatedness of all the proteins to be multiply aligned. In ClustalW, the tree is described with a written syntax rather than a graphical output (Figs. 10.2 and 10.4, stage 2). You can see that the smallest distance value occurs between human and mouse RBP. This tree can also be displayed graphically at the ClustalW site by using the JalView option

Stage 3: progressively align the sequences following the branching order of the tree

```
CLUSTAL W (1.82) multiple sequence alignment


gi|5803139|ref|NP_006735.1|          MKWVWALLLLAAWA--AAERD------CRVSSFR----VKENFDKARFSG 38
gi|12843160|dbj|BAB25881.1|          MEWVWALVLLAALGGGSAERD------CRVSSFR----VKENFDKARFSG 40
gi|4502163|ref|NP_001638.1|          --MVMLLLLLSALAGLFGAAEGQAFHLGKCPNPP----VQENFDVNKYLG 44
gi|732003|sp|P39281|BLC_ECOLI        ---MRLLPLVAAATAAFLVVA------CSSPTPPRGVTVVNNFDAKRYLG 41
gi|127528|sp|P11590|MUP4_MOUSE       ----MKLLLCLGLTLVCIHAE-------EATSKG------QNLNVEKING 33
                                           * *  .                  ..        :*::  :  *


gi|5803139|ref|NP_006735.1|          TWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCAD 88
gi|12843160|dbj|BAB25881.1|          LWYAIAKKDPEGLFLQDNIIAEFSVDEKGHMSATAKGRVRLLSNWEVCAD 90
gi|4502163|ref|NP_001638.1|          RWYEIEK-IPTTFENGRCIQANYSLMENGKIKVLNQELR--ADGTVNQIE 91
gi|732003|sp|P39281|BLC_ECOLI        TWYEIARFDHRFERGLEKVTATYSLRDDGGLNVINKGYNP-DRGMWQQSE 90
gi|127528|sp|P11590|MUP4_MOUSE       EWFSILLASDKREK-IEEHGSMRVFVEHIHVLENSLAFKFHTVIDGECSE 82
                                      *: :               :   . :   :                   :


gi|5803139|ref|NP_006735.1|          MVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRL 138
gi|12843160|dbj|BAB25881.1|          MVGTFTDTEDPAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFALQYSCRL 140
gi|4502163|ref|NP_001638.1|          GEATPVNLTEPAKLEVKFS---WFMP--SAPYWILATDYENYALVYSC-- 134
gi|732003|sp|P39281|BLC_ECOLI        GKAYFTGAPTRAALKVSFFG-------------PFYGGYNVIALDREYR- 126
gi|127528|sp|P11590|MUP4_MOUSE       IFLVADKTEKAGEYSVMYDG--------FNTFTILKTDYDNYIMFHLIN- 123
                                              .  .: :                .  .*:   :


gi|5803139|ref|NP_006735.1|          LNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNG 188
gi|12843160|dbj|BAB25881.1|          QNLDGTCADSYSFVFSRDPNGLSPETRRLVRQRQEELCLERQYRWIEHNG 190
gi|4502163|ref|NP_001638.1|          TCIIQLFHVDFAWILARNPN-LPPETVDSLKNILTSNNIDVKKMTVTDQV 183
gi|732003|sp|P39281|BLC_ECOLI        HALVCGPDRDYLWILSRTPT-ISDEVKQEMLAVATREGFDVSKFIWVQQP 175
gi|127528|sp|P11590|MUP4_MOUSE       --EKDGKTFQLMELYGRKADLNSDIKEKFVKLCEEHGIIKENIIDLTKTN 171
                                             .   : .* .   .      :        :  .     .


gi|5803139|ref|NP_006735.1|          YCDGRSERNLL 199
gi|12843160|dbj|BAB25881.1|          YCQSRPSRNSL 201
gi|4502163|ref|NP_001638.1|          NCPKLS----- 189
gi|732003|sp|P39281|BLC_ECOLI        GS--------- 177
gi|127528|sp|P11590|MUP4_MOUSE       RCLKARE---- 178
```

FIGURE 10.3. *Multiple sequence alignment of five distantly related lipocalins (see Fig. 10.2). The output is from ClustalW using the progressive alignment algorithm of Feng and Doolittle. In stage 3, a multiple sequence alignment is created by performing progressive sequence alignments. First, the two closest sequences are aligned (top two lines). Next, further sequences are added in an order based on their position in the guide tree. An asterisk indicates positions in which the amino acid residue is 100% conserved in a column; a colon indicates conservative substitutions; a dot indicates less conservative substitutions.*

(Figs. 10.2 and 10.4, stage 2). Such a tree is not considered a phylogenetic tree but instead is a guide tree that is used in the third stage to define the order in which sequences are added to a multiple alignment.

3. In stage 3, the multiple sequence alignment is created in a series of steps based on the order presented in the guide tree. The algorithm first selects the two most closely related sequences from the guide tree and creates a pairwise alignment. The next sequence is either added to the pairwise alignment (to generate an aligned group of three sequences) or used in another pairwise alignment. This procedure is continued progressively until a full alignment is obtained (Figs. 10.3 and 10.5, stage 3).

The Feng-Doolittle approach includes the rule "once a gap, always a gap." The most closely related pair of sequences is aligned first. As further sequences are added to the alignment, there are many ways that gaps could be included. The rationale for the "once a gap, always a gap" rule is that the two most closely related sequences that are initially aligned should be weighted most heavily in assigning gaps.

Stage 1: generate a series of pairwise sequence alignments

```
Sequence format is Pearson
Sequence 1: gi|5803139|ref|NP_006735.1|          199 aa
Sequence 2: gi|6174963|sp|Q00724|RETB_MOUS       201 aa
Sequence 3: gi|132407|sp|P04916|RETB_RAT         201 aa
Sequence 4: gi|89271|pir||A39486                 201 aa
Sequence 5: gi|132403|sp|P18902|RETB_BOVIN       183 aa


        Start of Pairwise alignments
        Aligning...
        Sequences (1:2) Aligned. Score:  84
        Sequences (1:3) Aligned. Score:  84
        Sequences (1:4) Aligned. Score:  91
        Sequences (1:5) Aligned. Score:  92
        Sequences (2:3) Aligned. Score:  99
        Sequences (2:4) Aligned. Score:  86
        Sequences (2:5) Aligned. Score:  85
        Sequences (3:4) Aligned. Score:  85
        Sequences (3:5) Aligned. Score:  84
        Sequences (4:5) Aligned. Score:  96
```

Stage 2: create a guide tree, calculated from the distance matrix

```
(
(
gi|5803139|ref|NP_006735.1|:0.04284,
(
gi|6174963|sp|Q00724|RETB_MOUS:0.00075,
gi|132407|sp|P04916|RETB_RAT:0.00423)
:0.10542)
:0.01900,
gi|89271|pir||A39486:0.01924,
gi|132403|sp|P18902|RETB_BOVIN:0.01902);
```

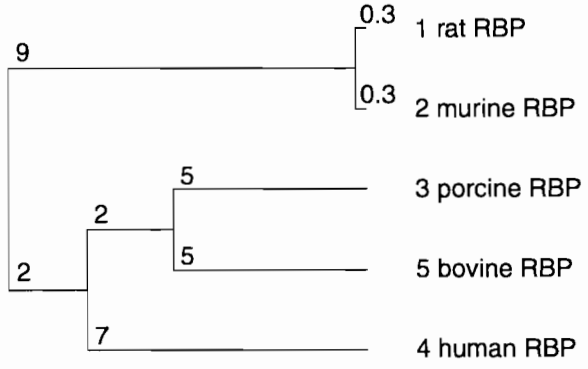


FIGURE 10.4. *Example of a multiple sequence alignment of closely related proteins using the progressive sequence aligment method of Feng and Doolittle as implemented by ClustalW. Compare these scores to those for distantly related proteins (Fig. 10.2), and note that the pairwise alignment scores are consistently higher and the distances (reflected in branch lengths on the guide tree) are much shorter.*

From Multiple Sequence Alignment to Profile Hidden Markov Models

The algorithms used to make progressive sequence alignments are closely related to dynamic programming, pairwise alignment approaches that we described in Chapter 3. Profile hidden Markov models (HMMs) provide a powerful application of alignment (Krogh, 1998; Eddy, 1998; Baldi et al., 1994; Birney, 2001). HMMs are probabilistic models that describe the likelihood that an amino acid residue occurs at each given position of an alignment. A profile HMM can convert a multiple sequence alignment into a position-specific scoring system. A common application of

```
Stage 3: progressively align the sequences following the branching order of the tree

CLUSTAL W (1.82) multiple sequence alignment

gi|89271|pir||A39486            MEWVWALVLLAALGSAQAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP 50
gi|132403|sp|P18902|RETB_BOVIN  ------------------ERDCRVSSFRVKENFDKARFAGTWYAMAKKDP 32
gi|5803139|ref|NP_006735.1|     MKWVWALLLLAAW--AAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDP 48
gi|6174963|sp|Q00724|RETB_MOUS  MEWVWALVLLAALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIAKKDP 50
gi|132407|sp|P04916|RETB_RAT    MEWVWALVLLAALGGGSAERDCRVSSFRVKENFDKARFSGLWYAIAKKDP 50
                                ********************:* ***:*****

gi|89271|pir||A39486            EGLFLQDNIVAEFSVDENGHMSATAKGRVRLLNNWDVCADMVGTFTDTED 100
gi|132403|sp|P18902|RETB_BOVIN  EGLFLQDNIVAEFSVDENGHMSATAKGRVRLLNNWDVCADMVGTFTDTED 82
gi|5803139|ref|NP_006735.1|     EGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTED 98
gi|6174963|sp|Q00724|RETB_MOUS  EGLFLQDNIIAEFSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTED 100
gi|132407|sp|P04916|RETB_RAT    EGLFLQDNIIAEFSVDEKGHMSATAKGRVRLLSNWEVCADMVGTFTDTED 100
                                *********:*******.*:************.**:**************

gi|89271|pir||A39486            PAKFKMKYWGVASFLQKGNDDHWIIDTDYDTYAAQYSCRLQNLDGTCADS 150
gi|132403|sp|P18902|RETB_BOVIN  PAKFKMKYWGVASFLQKGNDDHWIIDTDYETFAVQYSCRLLNLDGTCADS 132
gi|5803139|ref|NP_006735.1|     PAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADS 148
gi|6174963|sp|Q00724|RETB_MOUS  PAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFALQYSCRLQNLDGTCADS 150
gi|132407|sp|P04916|RETB_RAT    PAKFKMKYWGVASFLQRGNDDHWIIDTDYDTFALQYSCRLQNLDGTCADS 150
                                ***************:*******:****:*:* ****** ********

gi|89271|pir||A39486            YSFVFARDPHGFSPEVQKIVRQRQEELCLARQYRLITHNGYCDGKSERNI 200
gi|132403|sp|P18902|RETB_BOVIN  YSFVFARDPSGFSPEVQKIVRQRQEELCLARQYRLIPHNGYCDGKSERNI 182
gi|5803139|ref|NP_006735.1|     YSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNL 198
gi|6174963|sp|Q00724|RETB_MOUS  YSFVFSRDPNGLSPETRRLVRQRQEELCLERQYRWIEHNGYCQSRPSRNS 200
gi|132407|sp|P04916|RETB_RAT    YSFVFSRDPNGLTPETRRLVRQRQEELCLERQYRWIEHNGYCQSRPSRNS 200
                                *****:*** *:.**.:::********** **** * *****:.:..**

gi|89271|pir||A39486            L 201
gi|132403|sp|P18902|RETB_BOVIN  L 183
gi|5803139|ref|NP_006735.1|     L 199
gi|6174963|sp|Q00724|RETB_MOUS  L 201
gi|132407|sp|P04916|RETB_RAT    L 201
                                *
```

FIGURE 10.5. *Multiple sequence of five closely related retinol-binding protein orthologs (see Fig. 10.4). The output is from ClustalW using the progressive alignment algorithm of Feng and Doolittle.*

BOX 10-1
Similarity Versus Distance Measures

Trees that represent protein or nucleic acid sequences usually display the differences between various sequences. One way to measure distances is to count the number of mismatches in a pairwise alignment. Another method, employed by the Feng and Doolittle progressive alignment algorithm, is to convert similarity scores to distance scores. Similarity scores are calculated from a series of pairwise alignments among all the proteins being multiply aligned. The similarity scores S between two sequences (i, j) are converted to distance scores D using the equation

$$D = -\ln S_{\text{eff}} \tag{10.1}$$

where

$$S_{\text{eff}} = \frac{S_{\text{real}(ij)} - S_{\text{rand}(ij)}}{S_{\text{iden}(ij)} - S_{\text{rand}(ij)}} \times 100 \tag{10.2}$$

Here, $S_{\text{real}(ij)}$ describes the observed similarity score for two aligned sequences i and j, $S_{\text{iden}(ij)}$ is the average of the two scores for the two sequences compared to themselves (if score i compared to i receives a score of 20 and score j compared to j receives a score of 10, then $S_{\text{iden}(ij)} = 15$); $S_{\text{rand}(ij)}$ is the mean alignment score derived from many (e.g., 1000) random shufflings of the sequences; and S_{eff} is a normalized score. If sequences i,j have no similarity, then $S_{\text{eff}} = 0$ and the distance is infinite. If sequences i,j are identical, then $S_{\text{eff}} = 1$ and the distance is 0.

(*a*)

	12345
human RBP	GTWYA
mouse RBP	GLWYA
apolipoprotein D	GRWYE
E. Coli BLC	GTWYE
mouse MUP4	GEWFS

(*b*)

	position				
Probability	1	2	3	4	5
p(G)	1.0				
p(T)		0.4			
p(L)		0.2			
p(R)		0.2			
p(E)		0.2			0.4
p(W)			1.0		
p(Y)				0.8	
p(F)				0.2	
p(A)					0.4
p(S)					0.2

(*c*) p(GEWYE) = (1.0)(0.2)(1.0)(0.8)(0.4) = 0.064
Log odds score = ln(1.0) + ln(0.2) + ln(1.0) + ln(0.8) + ln (0.4) = -2.75

(*d*)

FIGURE 10.6. *Hidden Markov models describe alignments based on the probability of amino acids occurring in an aligned column. This is conceptually related to the position-specific scoring matrix used by PSI-BLAST. (a) An alignment of five lipocalins is shown in the region of the GXW motif (see Fig. 10.3). (b) The probability of each residue occurring in each aligned column of residues is calculated. (c) From these probabilities, a score is derived for any query such as GEWYE. Note that the actual score will also account for gaps and other parameters. Also note that this is a position-specific scoring scheme; for example, there is a different probability of the amino acid residue glutamate occurring in position 2 versus 5. (d) The probabilities associated with each position of the alignment can be displayed in boxes representing states.*

profile HMMs is the query of a single protein sequence of interest against a database of profile HMMs (as done in the Pfam database). Another application is to use a profile HMM as the query in a database search.

Consider the five amino acid residues in the conserved GXW region of five lipocalins (Fig. 10.6*a*). An HMM can be calculated by estimating the probability of occurrence of each amino acid in the five positions (Fig. 10.6*b*). In this sense, the HMM approach resembles the position-specific scoring matrix (PSSM) calculation of PSI-BLAST (Chapter 5). From the HMM probabilities, a score can be derived for the occurrence of any specific pattern of a related query, such as GEWYE (Fig. 10.6*c*). The HMM is a model that can be described in terms of "states" at each position of a sequence (Fig. 10.6*d*).

A profile HMM is more complex than a PSSM. It is constructed from an initial multiple sequence alignment to define a set of probabilities. The structure of a profile HMM is shown in Fig. 10.7 (Krogh et al., 1994; Krogh, 1998). Along the bottom row is a series of main states (from "begin" to m_1–m_5, then "end"). These states might correspond to residues of an amino acid sequence such as GTWYA. The second row consists of insert states (Fig. 10.7, diamond-shaped objects labeled i_1–i_5). These states model variable regions in the alignment, allowing sequences to be inserted as necessary. The third row, at the top, consists of circles called delete states. These correspond to gaps: They provide a path to skip a column (or columns) in the multiple sequence alignment.

Overall the protein sequence of an HMM is defined by a series of states that are connected to each other by state transitions (Fig. 10.7, arrows). Each state has

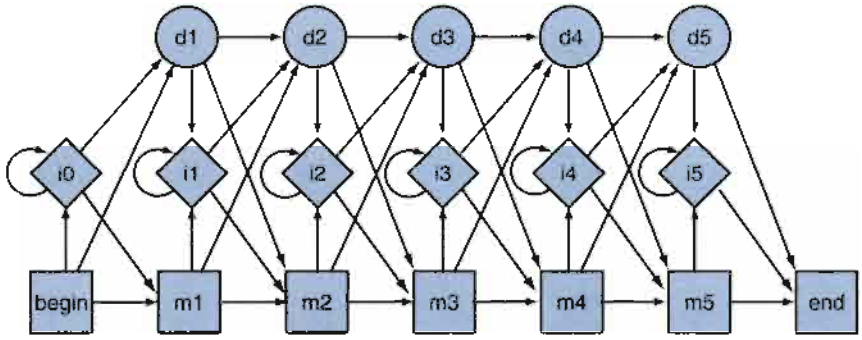


FIGURE 10.7. *The structure of a hidden Markov model. The HMM consists of a series of states associated with probabilities. The "main states" are shown in boxes along the bottom (from begin to end, with $m_1 - m_5$ in between). These main states model the columns of a multiple sequence alignment, and the probability distribution is the frequency of amino acids (see Fig. 10.6 d). The "insert states" are in diamond-shaped objects and represent insertions. For example, in a multiple sequence alignment some of the proteins might have an inserted region of amino acids, and these would be modeled by insert states. The "deletion states" ($d_1 - d_5$) represent gaps in the alignment.*

a "symbol emission" probability distribution for matching a particular amino acid residue. The symbol sequence of an HMM is an observed sequence that resembles a consensus for the multiple sequence alignment. There are also "state sequences" that describe the path followed along the Markov chain.

The HMM consists of both observed symbols (such as the amino acid residues in a sequence modeled by the HMM) and a hidden state sequence which is inferred probabilistically from the observed sequence. This is why the model is called "hidden."

Profile HMMs are important because they provide a powerful way to search databases for distantly related homologs. Thus HMM methods complement standard BLAST searching. Profile HMMs can define a protein or gene family, and databases of profile HMMs are searchable. Three of the most commonly used web-based databases of profile HMMs are Pfam, SMART, and TIGRFAMs (described below).

HMMs can be created using the HMMER program. You can build a profile HMM with the hmmbuild program, which reads a multiple sequence alignment as input (Fig. 10.8*a*). (The resulting profile HMM is a model that is global with respect to the HMM and local with respect to the sequences it matches in a database search.) Next, the hmmcalibrate program matches a set of 5000 random sequences to the profile HMM, fits the scores to an extreme-value distribution (EVD; Chapter 4), and calculates the EVD parameters that are necessary to estimate the statistical significance of database matches (Fig. 10.8*b*). The profile HMM can then be used to search a database using the hmmsearch program.

HMMER is available at ▸http://hmmer.wustl.edu. It was written by Sean Eddy. The program is designed to run on UNIX platforms.

When the profile HMM was built from a multiple sequence alignment of five highly divergent lipocalins and used to search the GenBank nonredundant database, there were many database matches (Fig. 10.9*a*), including RBP4 (Fig. 10.9*b*) and bacterial lipocalins (Fig. 10.9*c*). In contrast, when an alignment of five closely related RBP orthologs was used to generate a profile HMM, the output exclusively consisted of additional RBPs (Fig. 10.10*a*). No bacterial homologs were detected. The alignment to human RBP4 received a higher bit score and lower *E* value. The profile HMM is a model that is sensitive to the multiple sequence alignment used as input.

The two main uses of HMMER are to search a sequence database with a single profile HMM, as described above (Figs. 10.9 and 10.10), or to search a single query sequence against a library of HMMs. Pfam is an example of such a library.

(*a*)

```
root@localhost hmmer]# hmmbuild lipocalins.hmm x.msf
hmmbuild - build a hidden Markov model from an alignment
HMMER 2.2g (August 2001)
Copyright (C) 1992-2001 HHMI/Washington University School of Medicine
Freely distributed under the GNU General Public License (GPL)
---------------------------------------
Alignment:      #1
Number of sequences: 5
Number of columns:  234

Determining effective sequence number    ... done. [4]
Weighting sequences heuristically        ... done.
Constructing model architecture          ... done.
Converting counts to probabilities       ... done.
Setting model name, etc.                 ... done. [x]

Constructed a profile HMM (length 230)
Average score:   411.45 bits  <-------- 1
Minimum score:   353.73 bits
Maximum score:   460.63 bits
Std. deviation:  52.58 bits

Finalizing model configuration           ... done.
Saving model to file                     ... done.
//
```

(*b*)

```
root@localhost hmmer]# hmmcalibrate lipocalins.hmm
hmmcalibrate -- calibrate HMM search statistics
HMMER 2.2g (August 2001)
Copyright (C) 1992-2001 HHMI/Washington University School of Medicine
Freely distributed under the GNU General Public License (GPL)
---------------------------------------
HMM file:                 lipocalins.hmm
Length distribution mean: 325
Length distribution s.d.: 200
Number of samples:        5000
random seed:              1034351005
histogram(s) saved to:    [not saved]
POSIX threads:            2
---------------------------------

HMM   :x
mu    : -123.894508
lambda :  0.179608
max   : -79.334000
//
```

FIGURE 10.8. *The HMMER program can be used to create a profile HMM using a multiple sequence alignment as input. The program was obtained from* ▸ *http://hmmer.wustl.edu and downloaded on a machine running the Linux operating system. The same five lipocalins shown in Figures 10.1–10.3 were multiply aligned with GCG (see Fig. 10.13) and entered into HMMER. (a) A profile HMM was built with the hmmbuild program, then (b) calibrated against a database of 5000 sequences. Mu and lambda values are thus calculated to provide statistical evaluations of the HMM. The average score of the profile HMM was 411 bits [arrow 1 in (a)]. When a profile HMM was built using five closely related retinol-binding proteins, the average score was 600.89 ± 24 bits (not shown).*

TWO MULTIPLE SEQUENCE ALIGNMENT PROGRAMS

In describing web-based resources for studying protein families, we will examine two kinds of multiple sequence alignment programs:

1. Many databases of multiple sequence alignments are available. These may be searched using text (i.e., a keyword search) or using any query sequence. The query may be an already known sequence (such as RBP) or any novel protein

(*a*)

```
Scores for complete sequences (score includes all domains):
Sequence                                  Description        Score    E-value  N
--------                                  -----------        -----    ------- ---
gi|20888903|ref|XP_129259.1|              (XM_129259) ret   461.1   1.9e-133   1
gi|132407|sp|P04916|RETB_RAT              Plasma retinol-    458.0   1.7e-132   1
gi|20548126|ref|XP_005907.5|              (XM_005907) sim   454.9   1.4e-131   1
gi|5803139|ref|NP_006735.1|               (NM_006744) ret   454.6   1.7e-131   1
gi|20141667|sp|P02753|RETB_HUMAN          Plasma retinol-   451.1   1.9e-130   1
.
.
gi|16767588|ref|NP_463203.1|              (NC_003197) out   318.2    1.9e-90   1
```

(*b*)

```
gi|5803139|ref|NP_006735.1|: domain 1 of 1, from 1 to 195: score 454.6, E = 1.7e-131
                  *->mkwVMkLLLLaALagvfgaAErdAfsvgkCrvpsPPRGfrVkeNFDv
                     mkwV++LLLLaA  +  +aAErd      Crv+s    frVkeNFD+
  gi|5803139     1    MKWVWALLLLAA--W--AAAERD------CRVSS----FRVKENFDK 33

                     erylGtWYeIaKkDprFErGL1lqdkItAeySleEhGsMsataeGrirVL
                     +r++GtWY++aKkDp  E GL+lqd+I+Ae+S++E+G+Msata+Gr+r+L
  gi|5803139    34 ARFSGTWYAMAKKDP--E-GLFLQDNIVAEFSVDETGQMSATAKGRVRLL 80

                     eNkelcADkvGTvtqiEGeasevfLtadPaklklKyaGvaSflqpGfddy
                     +N+++cAD+vGT+t++E           dPak+k+Ky+GvaSflq+G+dd+
  gi|5803139    81 NNWDVCADMVGTFTDTE----------DPAKFKMKYWGVASFLQKGNDDH 120

                     wIlaTDYdYnYAlvYSCtpiirLinEkDGkcadsyswilsRdPnGLspEt
                     wI++TDYd +YA++YSC     rL+n +DG+cadsys+++sRdPnGL+pE+
  gi|5803139   121 WIVDTDYD-TYAVQYSC----RLLN-LDGTCADSYSFVFSRDPNGLPPEA 164

                     keklrkvlteegidvkkqyiwitqnnyCpkrse<-*
                     ++++r+  ++e++++++qy++i++n+yC++rse
  gi|5803139   165 QKIVRQ--RQEELCLARQYRLIVHNGYCDGRSE    195
```

(*c*)

```
gi|16767588|ref|NP_463203.1|: domain 1 of 1, from 1 to 177: score 318.2, E = 1.9e-90
                  *->mkwVMkLLLLaALagvfgaAErdAfsvgkCrvpsPPRGfrVkeNFDv
                        M+LL+ +A       a ++ Af+v++C++p+PP+G++V++NFD+
  gi|1676758      1    ----MRLLPVVA------AVTA-AFLVVACSSPTPPKGVTVVNNFDA 36

                     erylGtWYeIaKkDprFErGL1lqdkItAeySleEhGsMsataeGrirVL
                     +rylGtWYeIa+ D+rFErGL   + +tA+ySl++         +G+i+V+
  gi|1676758     37 KRYLGTWYEIARLDHRFERGL---EQVTATYSLRD--------DGGINVI 75

                     eNkelcADkvGTvtqiEGeasevfLtadPaklklKyaGvaSflqpGfddy
                      Nk++++D+  +++ +EG+a   ++t+ P +++lK+    Sf++p++++y
  gi|1676758     76 -NKGYNPDR-EMWQKTEGKA---YFTGSPNRAALKV----SFFGPFYGGY 116

                     wIlaTDYdYnYAlvYSCtpiirLinEkDGkcadsyswilsRdPnGLspEt
                     +++a+D++Y++Alv  C+p     + +D      y+wilsR+P+ Ls E+
  gi|1676758    117 NVIALDREYRHALV--CGP-----D-RD------YLWILSRTPT-LSEEI 151

                     keklrkvlteegidvkkqyiwitqnnyCpkrse<-*
                     k+++++v+t+eg+dv k  iw++q     +  +
  gi|1676758    152 KQQMLAVATREGFDVNK-LIWVKQ----SG--S    177
```

FIGURE 10.9. *(a) The HMMER output includes a variety of lipocalins, including bacterial members, when the HMM is trained with a set of five distantly related lipocalins. The alignments of the HMM to (a) human RBP4 and (b) a bacterial lipocalin are shown.*

(such as a new lipocalin you have identified). In some databases, the query sequence you provide is incorporated into the multiple sequence alignment of a particular precomputed protein family (see below).

2. Multiple sequence alignments can be generated by the manual input of a group of protein or nucleic acid sequences of interest. In this approach, a database is not searched, and the alignment is customized to the sequences of interest.

(*a*)

```
Scores for complete sequences (score includes all domains):
Sequence                              Description       Score    E-value  N
--------                              -----------       -----    ------- ---
gi|3041715|sp|P27485|RETB_PIG          Plasma retinol-   614.2   1.6e-179   1
gi|89271|pir||A39486                  plasma retinol-   613.9   1.9e-179   1
gi|20888903|ref|XP_129259.1|          (XM_129259) ret   608.8   6.8e-178   1
gi|132407|sp|P04916|RETB_RAT          Plasma retinol-   608.0   1.1e-177   1
gi|20548126|ref|XP_005907.5|          (XM_005907) sim   607.3   1.9e-177   1
gi|20141667|sp|P02753|RETB_HUMAN      Plasma retinol-   605.3   7.2e-177   1
gi|5803139|ref|NP_006735.1|           (NM_006744) ret   600.2   2.6e-175   1
```

(*b*)

```
gi|5803139|ref|NP_006735.1|: domain 1 of 1, from 1 to 199: score 600.2, E = 2.6e-175
                    *->meWvWaLvLLaalGgasaERDCRvssFRvKEnFDKARFsGtWYAiAK
                       m+WvWaL+LLaa+  a+aERDCRvssFRvKEnFDKARFsGtWYA+AK
  gi|5803139      1    MKWVWALLLLAAW--AAAERDCRVSSFRVKENFDKARFSGTWYAMAK 45

                    KDPEGLFLqDnivAEFsvDEkGhmsAtAKGRvRLLnnWdvCADmvGtFtD
                    KDPEGLFLqDnivAEFsvDE+G+msAtAKGRvRLLnnWdvCADmvGtFtD
  gi|5803139     46 KDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTD 95

                    tEDPAKFKmKYWGvAsFLqkGnDDHWiiDtDYdtfAvqYsCRLlnLDGtC
                    tEDPAKFKmKYWGvAsFLqkGnDDHWi+DtDYdt+AvqYsCRLlnLDGtC
  gi|5803139     96 TEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTC 145

                    ADsYsFvFsRDPnGlsPEvqkivRqRqEELCLaRqYRlieHnGYCdgrse
                    ADsYsFvFsRDPnGl+PE+qkivRqRqEELCLaRqYRli+HnGYCdgrse
  gi|5803139    146 ADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSE 195

                    RniL<-*
                    Rn+L
  gi|5803139    196 RNLL    199
```

FIGURE 10.10. *(a) When the HMMER program is trained with a multiple sequence alignment consisting of five closely related RBPs, the output of a database search consists primarily of additional RBP orthologs. (b) Note that the E values are much lower than for the HMM shown in Figure 10.9, and the bit score for the alignment of the HMM to human RBP4 has a higher score. However, this search fails to identify bacterial lipocalins or other distantly related proteins.*

Databases of Multiple Sequence Alignments

Several different groups have created large databases of protein families or domains (Table 10.1). There are several distinctions between these various resources.

Pfam: Protein Family Database of Profile HMMs

Pfam is one of the most comprehensive databases of protein families (Bateman et al., 2000, 2002). It is a compilation of both multiple sequence alignments and profile HMMs of protein families. The database can be searched using text (keywords or protein names) or by entering sequence data. Pfam's home is at The Wellcome Trust Sanger Institute (Hinxton, United Kingdom) mirrored in the United States, Sweden, and France.

Pfam consists of two databases. Pfam-A is a manually curated collection of protein families in the form of multiple sequence alignments and profile HMMs. For each family, Pfam provides four features: annotation, a seed alignment, a profile HMM, and a full alignment. The full alignment can be quite large; the top 20 Pfam families each contain over 2500 sequences in their full alignment. The seed alignments contain a smaller number of representative family members. In addition to the expertly curated Pfam-A, Pfam-B is automatically generated from the ProDom database (see below). The data in Pfam-B are not of as high quality as

TABLE 10-1 Databases of Multiple Sequence Alignments

Name	Description	URL
BLOCKS	HMM-like profile HMM library; ungapped	▸http://www.blocks.fhcrc.org/
CDD	Conserved domain database (Pfam plus SMART)	▸http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml
Interpro	A unified resource combining PROSITE, PRINTS, ProDom and Pfam, SMART, and TIGRfam	▸http://www.ebi.ac.uk/interpro/index.html
iProClass database	From the Protein Information Resource	▸http://pir.georgetown.edu/iproclass
MetaFAM	Proteins compiled from six databases	▸http://metafam.ahc.umn.edu/
Pfam	Profile HMM library	▸http://Pfam.wustl.edu/http://www.sanger.ac.uk/Software/Pfam/
PIR-ALN	Database of alignments	
PRINTS	Protein fingerprints from SwissProt/TrEMBL	▸http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
ProDom	Uses PSI-BLAST to cluster SWISS-PROT	▸http://prodes.toulouse.inra.fr/prodom/doc/prodom.html
PROSITE	A dictionary of protein motifs	▸http://www.expasy.ch/prosite/
SMART	Simple Modular Architecture Research Tool	▸http://smart.embl-heidelberg.de/
TIGRFAMs	HMM library of protein families	▸http://tigrblast.tigr.org/web-hmm/

Abbreviations: HMM, hidden Markov model; MSA, multiple sequence alignment.

There are four URLs for Pfam: ▸http://Pfam.wustl.edu/ (U.S.), ▸http://www.sanger.ac.uk/Software/Pfam/ (U.K.), ▸http://www.cgb.ki.se/Pfam (Sweden), and ▸http://pfam.jouy.inra.fr (France). Version 9.0 (May 2003) contains 5724 protein families. Currently, over 69% of the proteins in SWISS-PROT 40 and TrEMBL-18 have at least one match to a Pfam-A family.

You can also search Pfam with a DNA query. Go to ▸http://www.sanger.ac.uk/Software/Pfam/dnasearch.shtml.

Pfam-A, nor are they annotated as completely. But Pfam-B serves as a useful supplement that makes the entire database more comprehensive.

We can see the main features of Pfam in a search for lipocalins. There are three main ways to access the database: by browsing for families, by entering a protein sequence search (with a protein accession number or sequence), and by entering a text search. From the front page, select a text-based search and enter "lipocalin" (Fig. 10.11). The output has four sections (Fig. 10.12). At the top, the annotation includes links to lipocalin entries in other databases. Next, the alignment can be retrieved in several formats. A portion of the multiple sequence format (MSF) output is shown in Figure 10.13. Another versatile output format is JalView. After selecting this option, press the JalView button (Fig. 10.14). A Java applet allows the multiple sequence alignment to be viewed, analyzed, and saved in a variety of ways (Fig. 10.15). The applet will display a principal components analysis (PCA) on the aligned family (Fig. 10.16). We described PCA as a technique to reduce highly dimensional data into two- (or three-) dimensional space (Fig. 7.16). Here, each protein in a multiple sequence alignment is represented as a point in space based on a distance metric, and outliers are easily identified. Similar information can be represented with a tree (Fig. 10.17) using the Java applet.

SMART

SMART (▸http://smart.embl-heidelberg.de/) currently has about 650 profile HMMs (as of October 2002).

The Simple Modular Architecture Research Tool (SMART) is a database of protein families implicated in cellular signaling, extracellular domains, and chromatin function (Schultz et al., 1998; Ponting et al., 1999). Like Pfam, SMART employs profile HMMs using HMMER software.

FIGURE 10.11. *The Pfam database allows text-based searches. See* ▸*http://pfam.wustl.edu/.*

Also like Pfam, the SMART database is searchable by sequence or by keyword (Fig. 10.18). A search with human RBP4 results in a description of the domain organization and composition for lipocalins (Fig. 10.19). Domains identified in a SMART search are extensively annotated with information on functional class, tertiary structure, and taxonomy. A search of the proteins encoded by the yeast genome revealed that 6.7% of its genes contained one or more signaling domains, more than the number detected by SwissProt or Pfam (Schultz et al., 1998).

Conserved Domain Database

The Conserved Domain Database (CDD) is an NCBI tool that allows sequence-based or text-based queries of Pfam and SMART. CDD uses reverse position-specific BLAST (RPS-BLAST) by comparing a query sequence to a set of many position-specific scoring matrices (PSSMs). RPS-BLAST is related to PSI-BLAST (Chapter 5) but is distinct because it searches against profiles generated from preselected alignments. The main purpose of CDD (and RPS-BLAST) is to identify conserved domains in the query sequence.

CDD is available at ▸http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml or through the main BLAST page (▸http://www.ncbi.nlm.nih.gov/BLAST/).

The result of a CDD search using HIV-1 pol as a query is shown in Figure 10.20. Note that both scores and *E* values are reported. For each PSSM match (i.e., for each motif that is identified), the query sequence is incorporated into a multiple alignment with that family. Also, as shown in Figure 10.20, additional sequences may be added to the multiple sequence alignment, including very similar or very diverse members of a family.

CDD can also be searched by entering a protein query sequence into the Domain Architecture Retrieval Tool (DART) at NCBI. The result of a DART search using HIV-1 pol is shown in Figure 10.21.

DART is available at ▸http://www.ncbi.nlm.nih.gov/Structure/lexington/html/overview.html

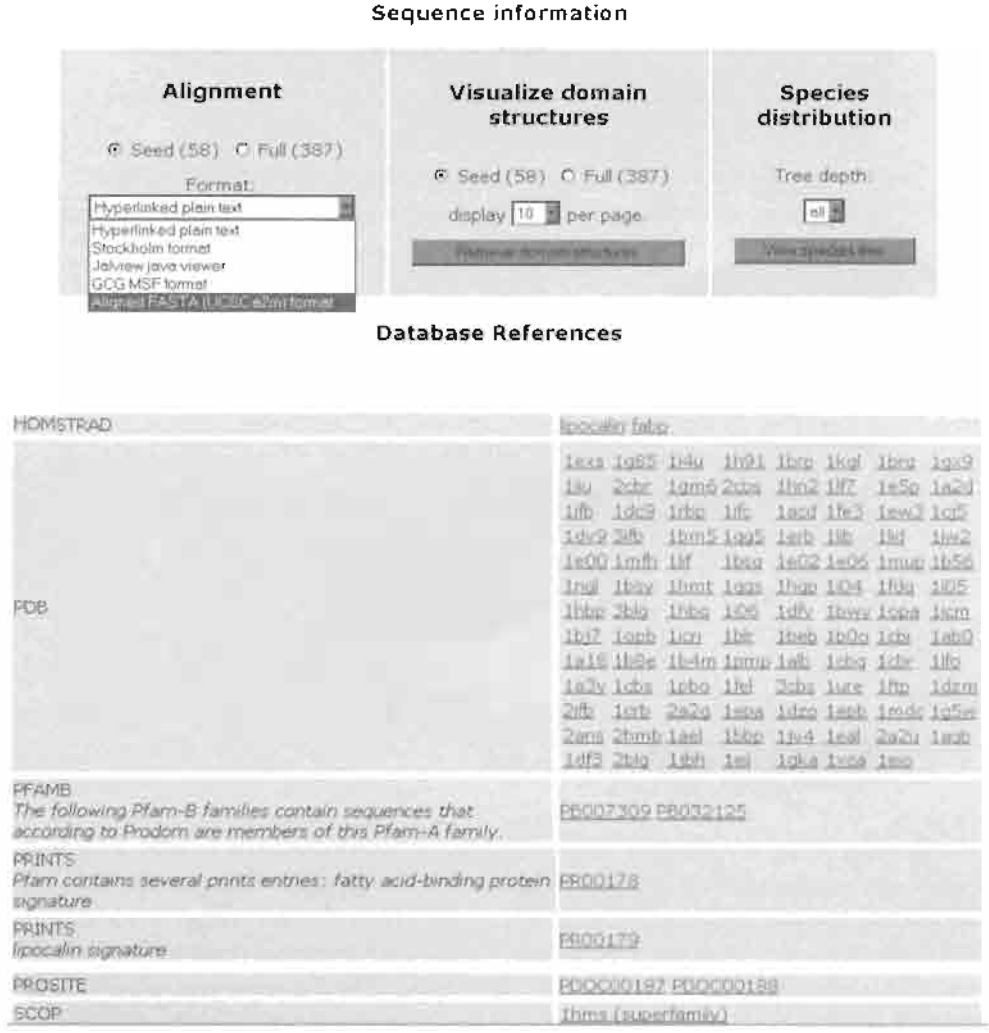


FIGURE 10.12. *A Pfam text search for "lipocalin" results in the identification of the Lipocalin/cytosolic fatty acid–binding protein family. A portion of that output is shown here. The accession number is PF00061, and the author is Sean Eddy, creator of Pfam. Links are provided to other databases of protein families and motifs. For clarity, not all the database references are shown. Pfam alignments can be shown in a variety of formats. The domain structure can be viewed as well as the profile HMM.*

BLOCKS

The BLOCKS database contains only ungapped multiple sequence alignments, corresponding to the most highly conserved regions of proteins (Henikoff et al., 1998, 1999). Blocks+ has been constructed by compiling five other databases, including PROSITE, Prints, ProDom, Domo, and Pfam (Table 10.1).

Currently, the BLOCKS database (▸http://www.blocks.fhcrc.org/) has 11,853 blocks from 2608 groups (Version 13.0).

PRINTS

The PRINTS database consists of protein "fingerprints" that define families in the SwissProt/TrEMBL databases (Attwood, 1999). A hyperlink to PRINTS outputs is the Colour Interactive Editor for Multiple Alignments (CINEMA) editor (Parry-Smith et al., 1998). This is a Java applet application that is integrated with software for analysis of the alignments.

Version 31 of PRINTS (2001) has 1550 database entries covering a total of 9531 motifs.

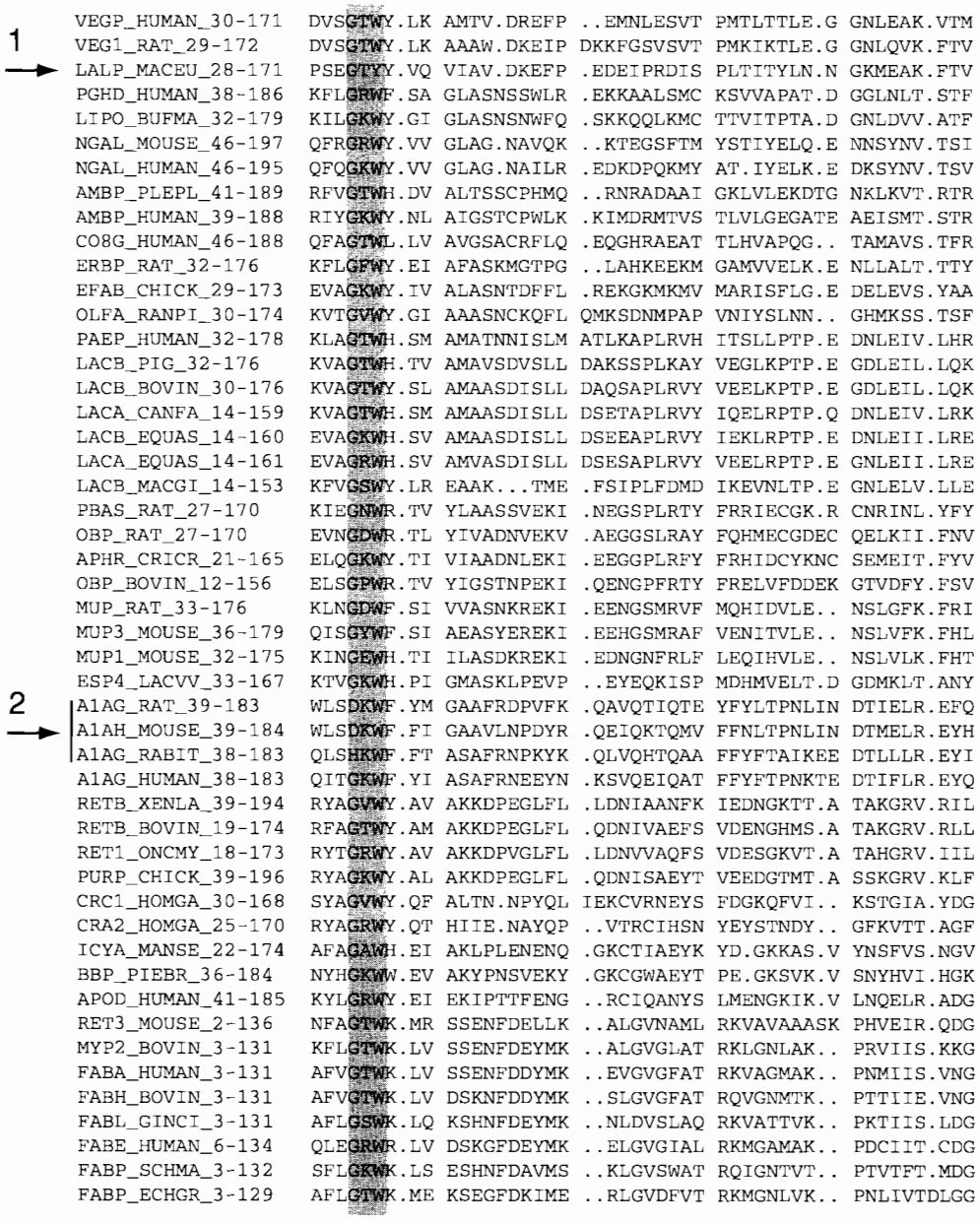
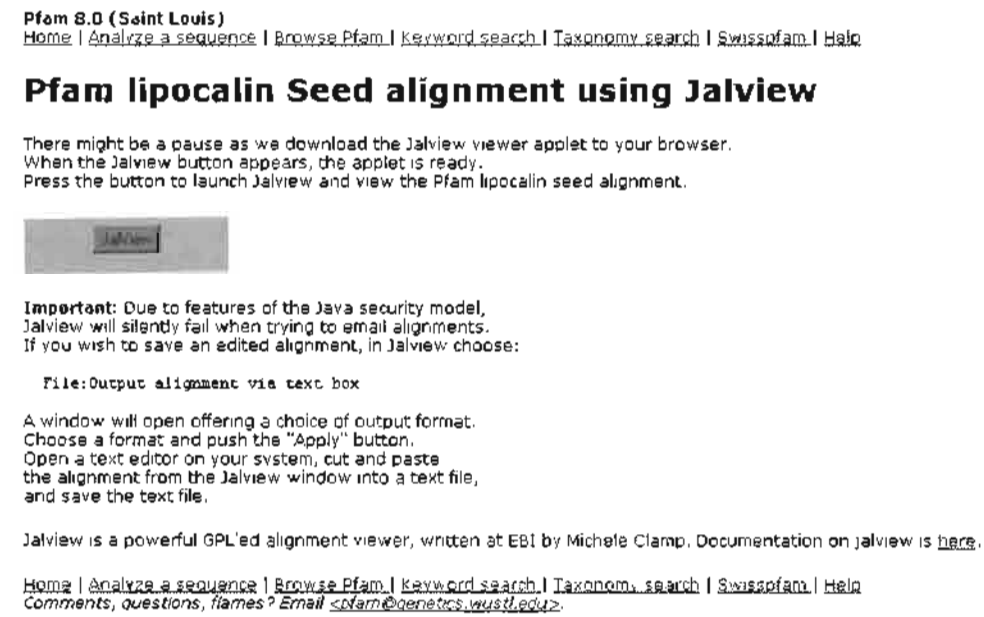


FIGURE 10.13. *Pfam alignments can be viewed in the GCG MSF format (multiple sequence format). A portion of the alignment is shown. Note that the GXW motif is well conserved (shaded residues), except for LALP_MACEU (late lactation protein from the tammar wallaby* Macropus eugenii; *P20462; arrow 1) and α 1-acid glycoproteins from rat, mouse, and rabbit (arrow 2; P02764, P07361, and P25227).*

PROSITE

The PROSITE database contains descriptions of protein domains and protein families (Hofmann et al., 1999; Sigrist et al., 2002). The main site page includes options for text-based searching (Fig. 10.22). A search for lipocalin results in a link to PDOC00187, the accession number for the document describing that family. We encountered a PROSITE document in Figure 8.8.

In its current release (June 2003) PROSITE contains descriptions of over 1000 families. It is accessed at ▸http://www.expasy.org/prosite/.



Pfam 8.0 (Saint Louis)
Home | Analyze a sequence | Browse Pfam | Keyword search | Taxonomy search | Swissptam | Help

Pfam lipocalin Seed alignment using Jalview

There might be a pause as we download the Jalview viewer applet to your browser.
When the Jalview button appears, the applet is ready.
Press the button to launch Jalview and view the Pfam lipocalin seed alignment.

Jalview

Important: Due to features of the Java security model,
Jalview will silently fail when trying to email alignments.
If you wish to save an edited alignment, in Jalview choose:

File:Output alignment via text box

A window will open offering a choice of output format.
Choose a format and push the "Apply" button.
Open a text editor on your system, cut and paste
the alignment from the Jalview window into a text file,
and save the text file.

Jalview is a powerful GPL'ed alignment viewer, written at EBI by Michele Clamp. Documentation on jalview is here.

Home | Analyze a sequence | Browse Pfam | Keyword search | Taxonomy search | Swissptam | Help
Comments, questions, flames? Email <pfam@genetics.wustl.edu>.

FIGURE 10.14. *A Pfam alignment can be retrieved in the JalView Java viewer format (see Fig. 10.12). When the rectangular JalView box appears, press it.*

Integrated Multiple Sequence Alignment Resources: InterPro, MetaFam, iProClass

A main theme of multiple sequence alignment databases is that while each employs a unique algorithm and search format, they are well integrated with each other. Another important idea is that individual databases such as Pfam and PROSITE have

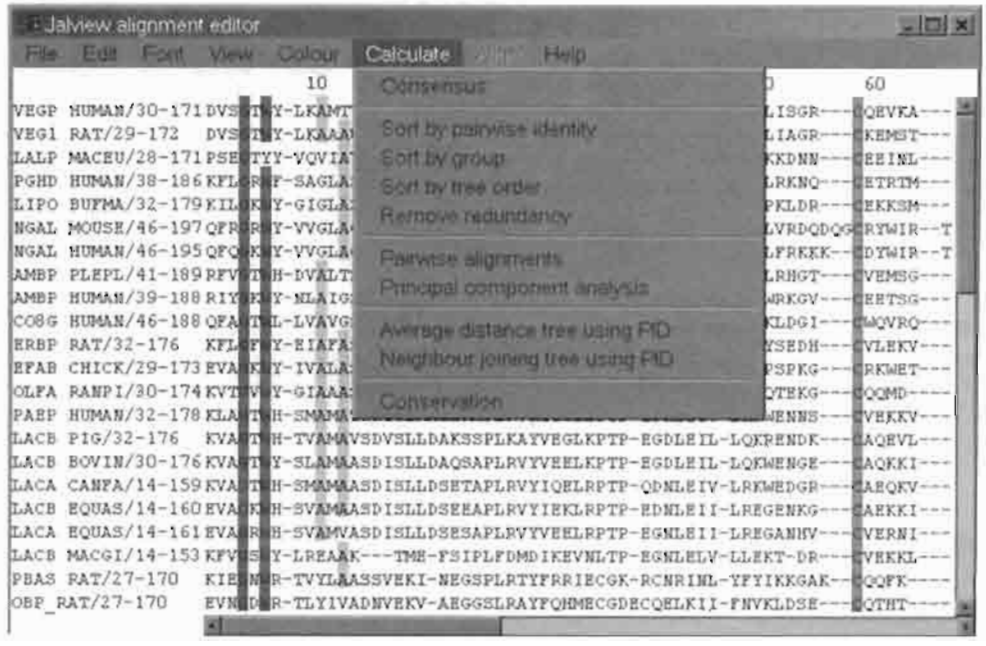


FIGURE 10.15. *The Pfam JalView applet displays a multiple sequence alignment of any Pfam protein family. The aligned residues can be viewed with a variety of color schemes. The relationships of the proteins within the family can be explored using a variety of algorithms, including principal components analysis (Fig. 10.16) and trees (Fig. 10.17).*

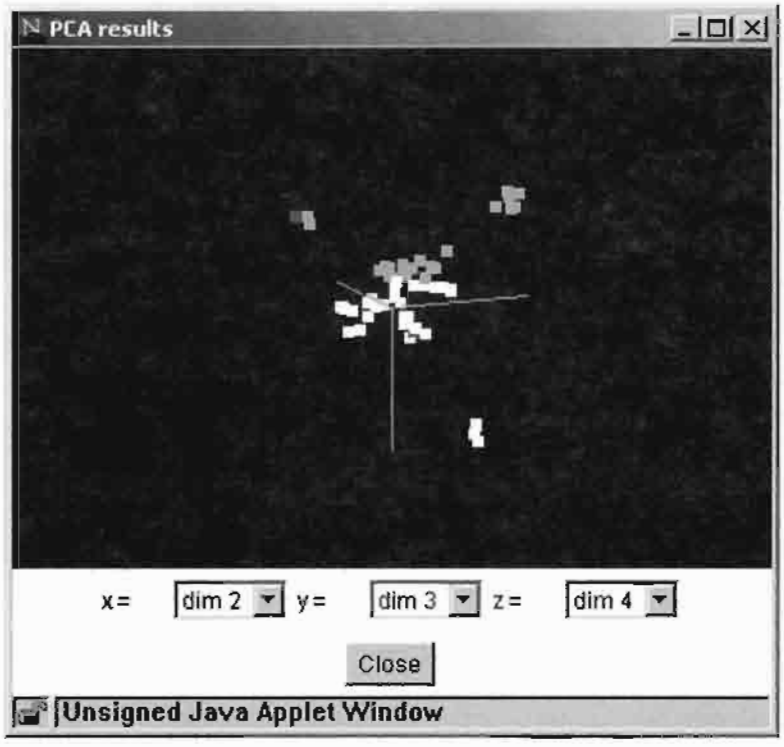


FIGURE 10.16. *A Pfam family analyzed with JalView can be visualized using PCA, a technique that represents each protein in the family as a point in space. The axes reflect distance between proteins and are units of percent variance. The axes may be rotated by mouse control. Several lipocalins are outliers (upper left, upper right, lower right). By selecting any protein, it is highlighted on the JalView multiple sequence alignment and can thus be identified.*

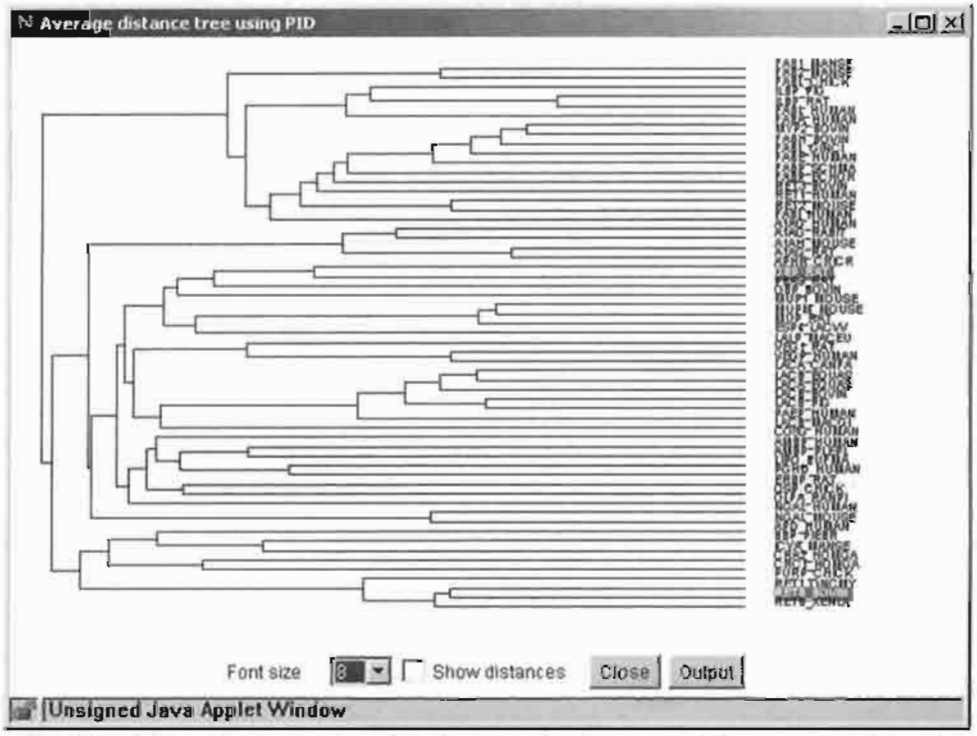


FIGURE 10.17. *JalView outputs from Pfam families include trees, such as this average distance tree of 58 lipocalins. (Phylogenetic trees will be described in Chapter 11.)*

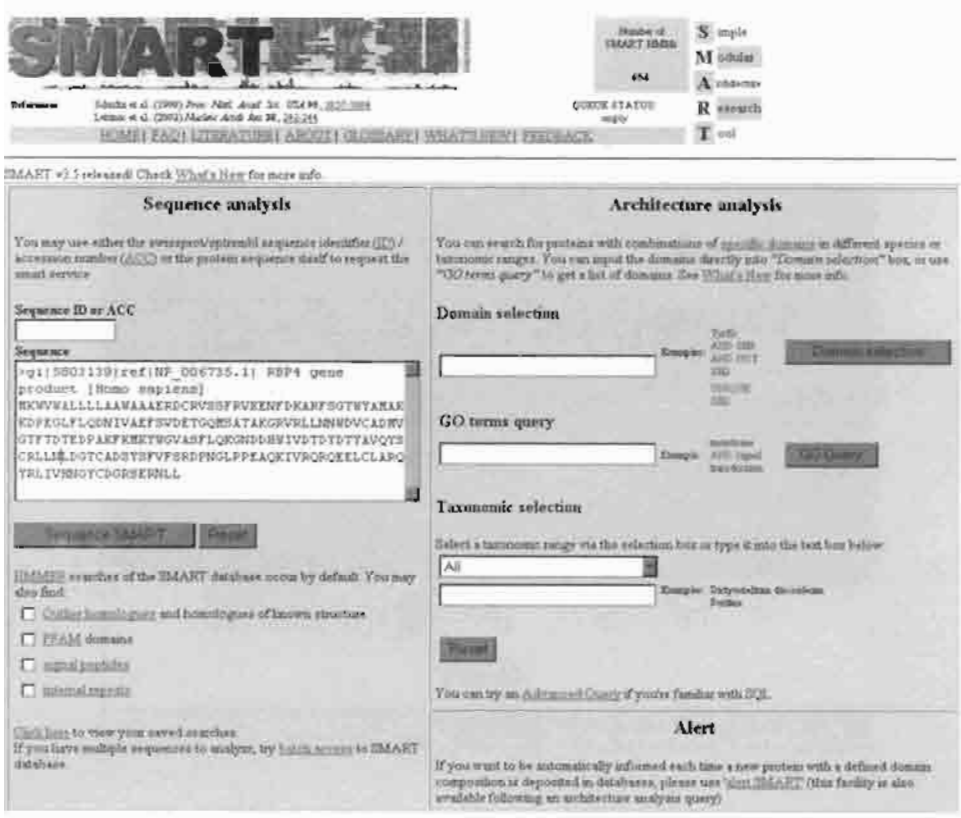


FIGURE 10.18. The SMART database (http://smart.emblheidelberg.de/) allows searches of protein sequences for conserved domains. Domains may be searched using a keyword search (at bottom), by analysis of domains (at right), or using a protein query (top left; the human RBP4 protein sequence is pasted into the query box in the FASTA format).

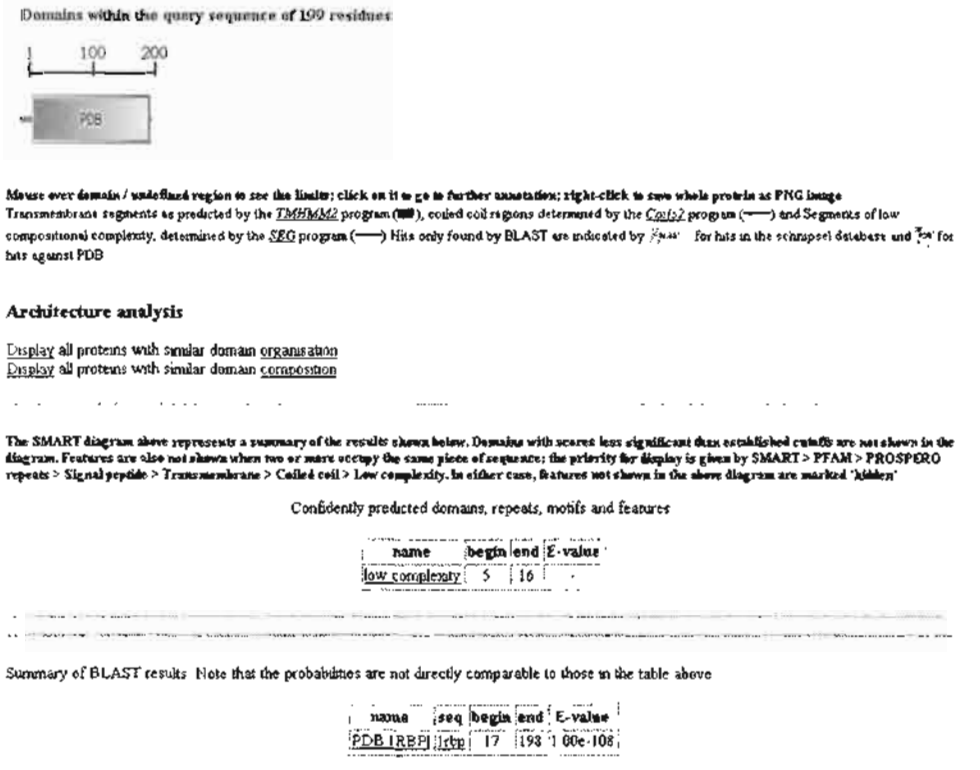
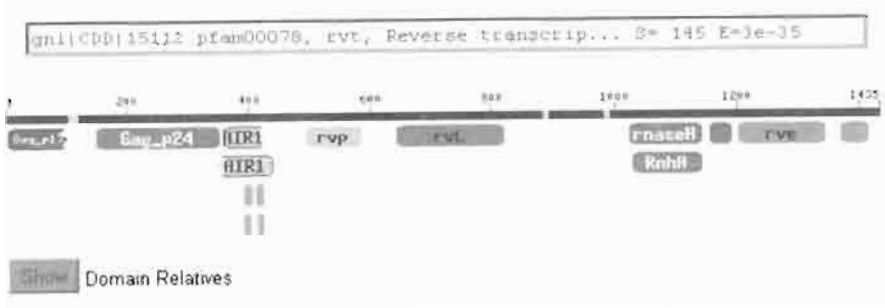


FIGURE 10.19. Result of a SMART search using human RBP4 as a query. The result shows a link to the PDB (Protein Data Bank) entry for RBP. A low-complexity region is identified, corresponding to the signal peptide motif that is characteristic of secreted proteins.

```
Query= gi|28872819|ref|NP_057849.4| Gag-Pol; Gag-Pol polyprotein
[Human immunodeficiency virus 1]
         (1435 letters)

Database: cdd.v1.61
           10,927 PSSMs; 2,688,589 total columns
```



● .. This CD alignment includes 3D structure To display structure, download Cn3D!

PSSMs producing significant alignments	Score (bits)	E value
● gnl|CDD|7686 pfam00607, Gag_p24, gag gene protein p24 (core nucleocapsid pr..	236	1e-62
● gnl|CDD|1099 pfam00540, Gag_p17, gag gene protein p17 (matrix protein) The.	187	7e-48
● gnl|CDD|15112 pfam00078, rvt, Reverse transcriptase (RNA-dependent DNA polym.	145	3e-35
● gnl|CDD|15111 pfam00075, rnaseH, RNase H. RNase H digests the RNA strand of..	139	3e-33
● gnl|CDD|9200 pfam00665, rve, Integrase core domain Integrase mediates inte.	115	4e-26
● gnl|CDD|5857 pfam00077, rvp, Retroviral aspartyl protease. Single domain as.	112	3e-25
● gnl|CDD|9376 pfam02022, Integrase_Zn, Integrase Zinc binding domain. Integr...	70.3	2e-12
● gnl|CDD|10202 COG0328, RnhA, Ribonuclease HI [DNA replication, recombination..	58.8	5e-09
● gnl|CDD|9178 pfam00552, integrase, Integrase DNA binding domain. Integrase..	51.5	7e-07
gnl|CDD|14211 COG5082, AIR1, Arginine methyltransferase-interacting protein,...	45.1	6e-05
gnl|CDD|14211 COG5082, AIR1, Arginine methyltransferase-interacting protein,	43.9	1e-04
● gnl|CDD|9059 pfam00098, zf-CCHC, Zinc knuckle. The zinc knuckle is a zinc b..	43.8	1e-04
● gnl|CDD|9059 pfam00098, zf-CCHC, Zinc knuckle The zinc knuckle is a zinc b..	41.5	7e-04
● gnl|CDD|8971 smart00343, ZnF_C2HC, zinc finger;	40.7	0.001
● gnl|CDD|8971 smart00343, ZnF_C2HC, zinc finger;	39.2	0.004

FIGURE 10.20. *Portion of the output from a CDD search from NCBI (▸http://www.ncbi.nlm.nih.gov/BLAST/) using RPS-BLAST. The query was HIV-1 pol (NP.057849), and the database matches are to Pfam families.*

evolved specific approaches to the problem of protein classification and analysis. Some databases employ HMMs; some focus on protein domains, while others assess smaller motifs. Integrated resources allow you to explore the features of a protein using several related algorithms in parallel.

At least three comprehensive resources have been developed to integrate most of the major alignment databases. The InterPro database provides an integration of PROSITE, PRINTS, ProDom, Pfam, and TIGRFAMs with cross-references to BLOCKS (Table 10.2) (Mulder et al., 2002, 2003). The project is coordinated by eight centers including EBI and The Wellcome Trust Sanger Institute (Apweiler et al., 2001).

InterPro is available at ▸http://www.ebi.ac.uk/interpro/. The September 2002 issue of the journal *Briefings in Bioinformatics* (volume 3, number 3) is devoted to InterPro and its member databases. Eighty-five percent of SwissProt and 75% of TrEMBL protein sequences currently have one or more hits in InterPro.

Another integrative resources is MetaFam (Shoop et al., 2001; Silverstein et al., 2001a,b). MetaFam is a unified web server that includes the alignments from 10 databases (Blocks+, DOMO, Pfam, PIR-ALN, PRINTS, PROSITE, ProDom, PROTOMAP, SBASE, and SYSTERS) (Silverstein, 2001b).

MetaFam is available at ▸http://metafam.ahc.umn.edu/.

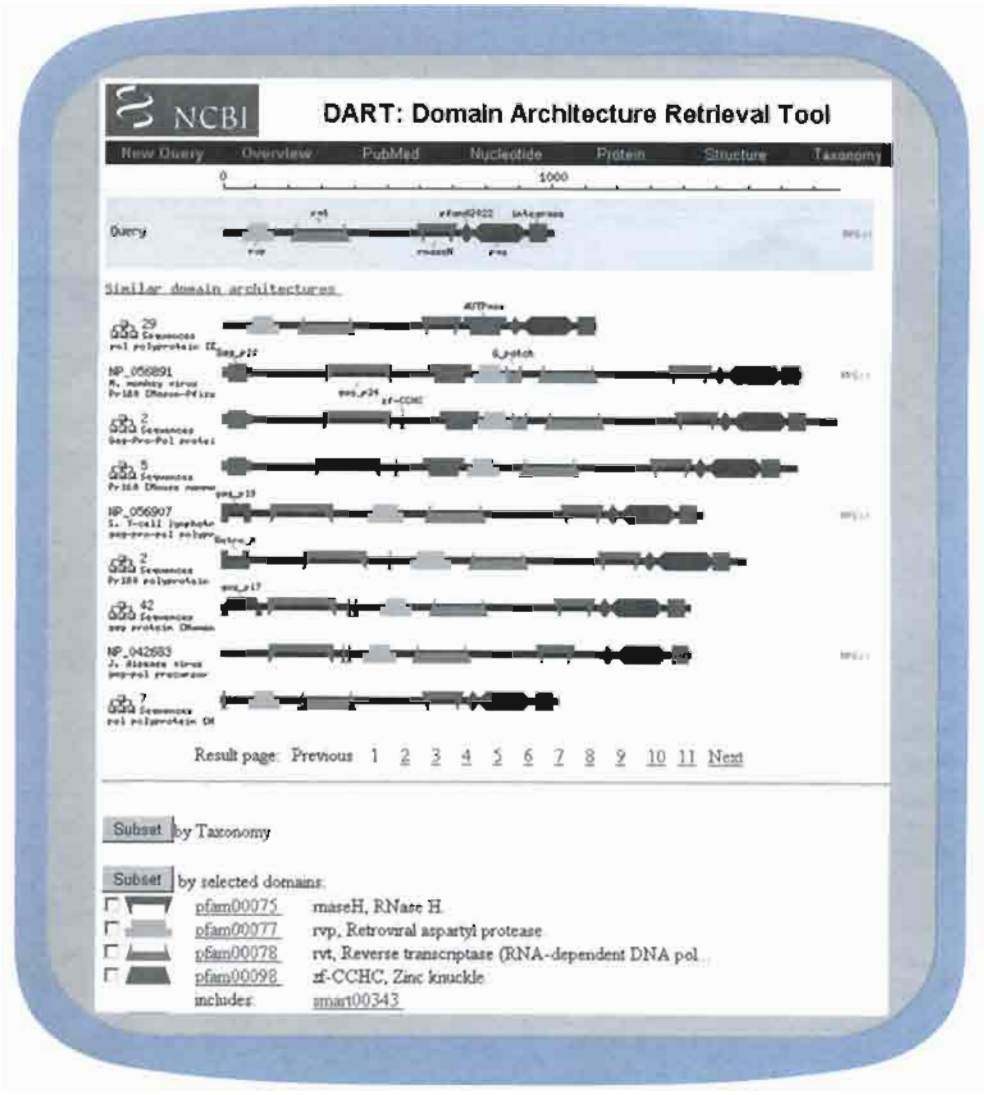


FIGURE 10.21. *DART is an NCBI tool that offers CDD searches. You can access DART via the CDD site of the NCBI structure page (▸http://www.ncbi.nlm.nih.gov/Structure/).*

You can access iProClass at ▸http://pir.georgetown.edu/iproclass/.

The iProClass organizes 200,000 nonredundant Protein Information Resource (PIR) and SwissProt proteins in 28,000 superfamilies, 2600 domains, 1300 motifs, and 280 posttranslational modification sites (Wu et al., 2001). iProClass has links to 30 other databases. Resources such as MetaFam, iProClass, and InterPro can be useful to identify conflicts between a variety of databases and to define the size of protein families.

PopSet

You can access PopSet via ▸http://www.ncbi.nlm.nih.gov/Entrez/.

We turn now to a specialized resource for multiple sequence alignments. PopSet (Population Data Study Sets) is a collection of aligned protein or DNA sequences within the NCBI Entrez site. These sequences are derived from studies of closely related sequences from population, phylogenetic, or mutation studies. In PopSet, enter the search "hiv-1 pol" and there are about 300 matches. Select one of these by Hermankova et al., and the output is a multiple sequence alignment of *pol* DNA isolated from a variety of patients (Fig. 10.23).

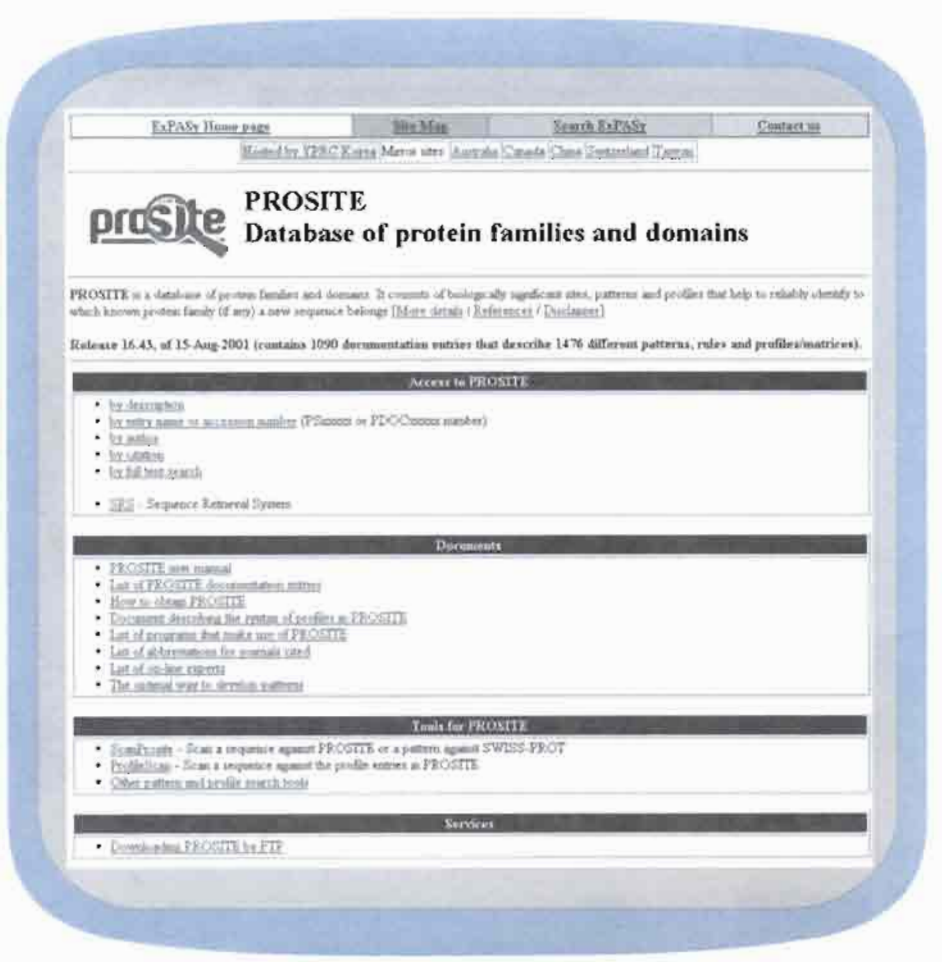


FIGURE 10.22. *The PROSITE database of protein families (http://www.expasy.ch/prosite/) is part of the Expert Protein Analysis System (ExPASy). PROSITE is a database of protein families, domain, and patterns. Created by Amos Bairoch and colleagues, PROSITE contains over 1000 families.*

Multiple Sequence Alignment Database Curation: Manual Versus Automated

Some databases are curated manually. This requires expert annotation; Sean Eddy and colleagues have curated Pfam, while Amos Bairoch and colleagues have curated PROSITE. BLOCKS and PRINTS are also manually annotated. Expert annotation is obviously difficult but has the great advantage of allowing judgments to be made on the protein family members. Programs such as DOMO, ProDom, and MetaFam use automated annotation. Errors in the alignment or the addition of unrelated sequences can be problematic, as discussed for PSI-BLAST (Chapter 5). However, automated annotation is valuable for exhaustive analyses of large data sets such as the thousands of predicted protein sequences derived from genome-sequencing projects.

TABLE 10-2 Databases on Which InterPro (Release 3.2) Is Based

Database	Contents
Pfam 7.3	3865 domains
PRINTS 33.0	1650 fingerprints
PROSITE 17.5	1565 profiles
ProDom 2001.3	1346 domains
SMART 3.1	509 domains
SwissProt 40.27	113,470 entries
TIGRFAMs 1.2	814 families
TrEMBL 21.12	685,610 entries

Source: From http://www.ebi.ac.uk/interpro/release_notes.html, October 2002.

User-Generated Multiple Sequence Alignments

The programs we have described so far all contain databases of protein families arranged in multiple sequence alignments. While these resources are very useful, it is often important to create your own alignment. The web sites and programs given in Table 10.3 allow you to generate a multiple sequence alignment using a specific set of sequences.

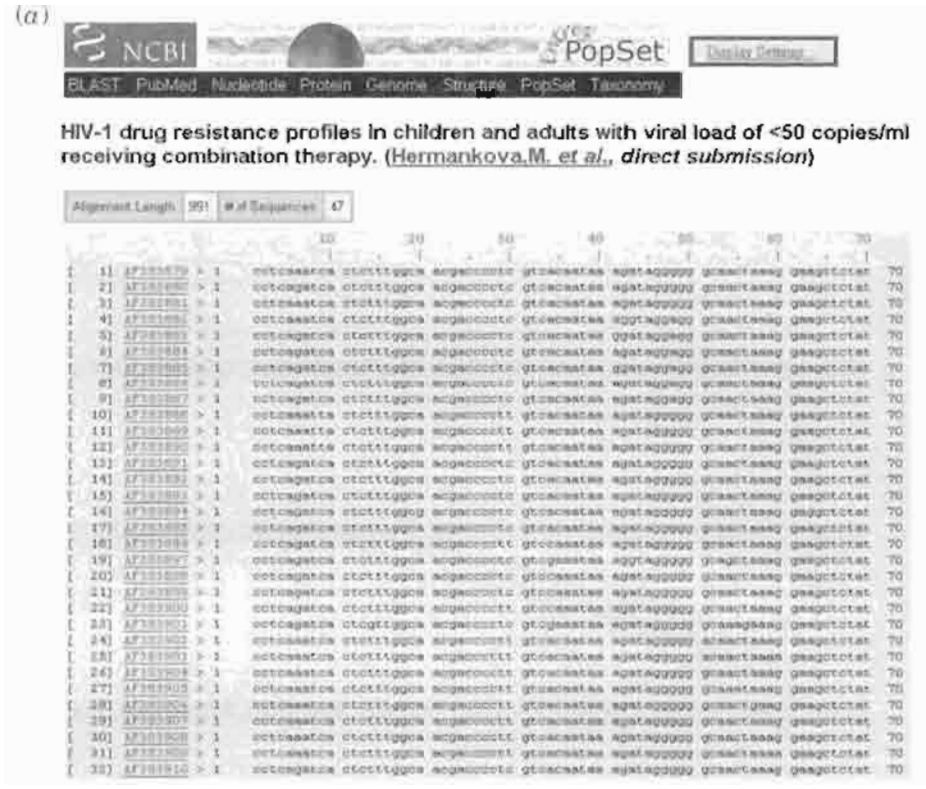


FIGURE 10.23. *A search of the PopSet database at NCBI retrieves multiple sequence alignments such as this one. The display settings box allows the alignment to be saved in a variety of formats, including FASTA and PHYLIP. This is useful for further phylogenetic studies of the alignment (Chapter 11). The display options include (a) the full sequences or (b) the variations.*

TABLE 10-3 Multiple Sequence Alignment Programs Available on World Wide Web

Program	Description	URL
AMAS (Analyse Multiply Aligned Sequences)	At the European Bioinformatics Institute; used to analyze premade MSAs	http://barton.ebi.ac.uk/servers/amas_server.html
CINEMA	Colour INteractive Editor for Multiple Alignments	http://www.bioinf.man.ac.uk/ dbbrowser/CINEMA2.1/
ClustalW	At the European Bioinformatics Institute and other sites	http://www2.ebi.ac.uk/clustalw/
ClustalX	Download by FTP	ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX/
DIALIGN	Especially useful for local MSA; from the University of Bielefeld, Germany	http://bibiserv.techfak.uni-bielefeld.de/dialign/
Match-Box Web Server 1.3	From the University of Namur, Belgium	http://www.fundp.ac.be/sciences/biologie/bms/matchbox_submit.shtml
MultAlin	From INRA (http://www.inra.fr/), Toulouse	http://prodes.toulouse.inra.fr/multalin/multalin.html
Multiple Sequence Alignment version 2.0	At the GeneStream server of the Institut de Génétique Humaine (http://www2.igh.cnrs.fr/home.eng.html)	http://xylian.igh.cnrs.fr/msa/msa.html
Multiple Sequence Alignment version 2.1	From Washington University, St. Louis	http://stateslab.wustl.edu/ibc/msa.html
Musca	From the IBM Bioinformatics Group	http://cbcsrv.watson.ibm.com/Tmsa.html
PileUp	Available through the SeqWeb or UNIX versions of the Genetics Computer Group (GCG)	http://www.gcg.com
T-COFFEE	Slower but more accurate than ClustalW for distantly related proteins	http://www.ch.embnet.org/software/TCoffee.html

Note: Additional algorithms are listed at ExPASy (http://kr.expasy.org/tools/#align).
Abbreviation: MSA, multiple sequence alignment.

ClustalW and ClustalX

ClustalW is a premier web-based tool for generating multiple sequence alignments (Thompson et al., 1994). The website includes a box for the input of sequences in the FASTA format (Fig. 10.1). (Other accepted formats include NBRF/PIR, EMBL/SwissProt, and GCG/MSF.) Earlier we saw a typical output of ClustalW (Fig. 10.2–10.5).

The ClustalX program is a windows interface for ClustalW (Thompson et al., 1997). The program is versatile, powerful, and easy to use, and it has been compiled for download onto a variety of platforms (see Table 10.3 for FTP site). After saving a group of lipocalins in the FASTA format as a text file, these sequences can be loaded into ClustalX (Fig. 10.24*a*). No gaps are included initially. The "Do complete alignment" command results in the generation of a multiple sequence alignment (Fig. 10.24*b*). This alignment should be visually inspected and modified as necessary. This often involves making informed judgments about the location of gaps and other parameters. Note that in Figure 10.24*b* there are two adjacent regions of gaps (arrows). The residues GQAFHL in the third protein may be thought of as an insertion with a gap in the other four proteins at that position; similarly, the residues RGVT could be considered an insertion in the fourth protein. One

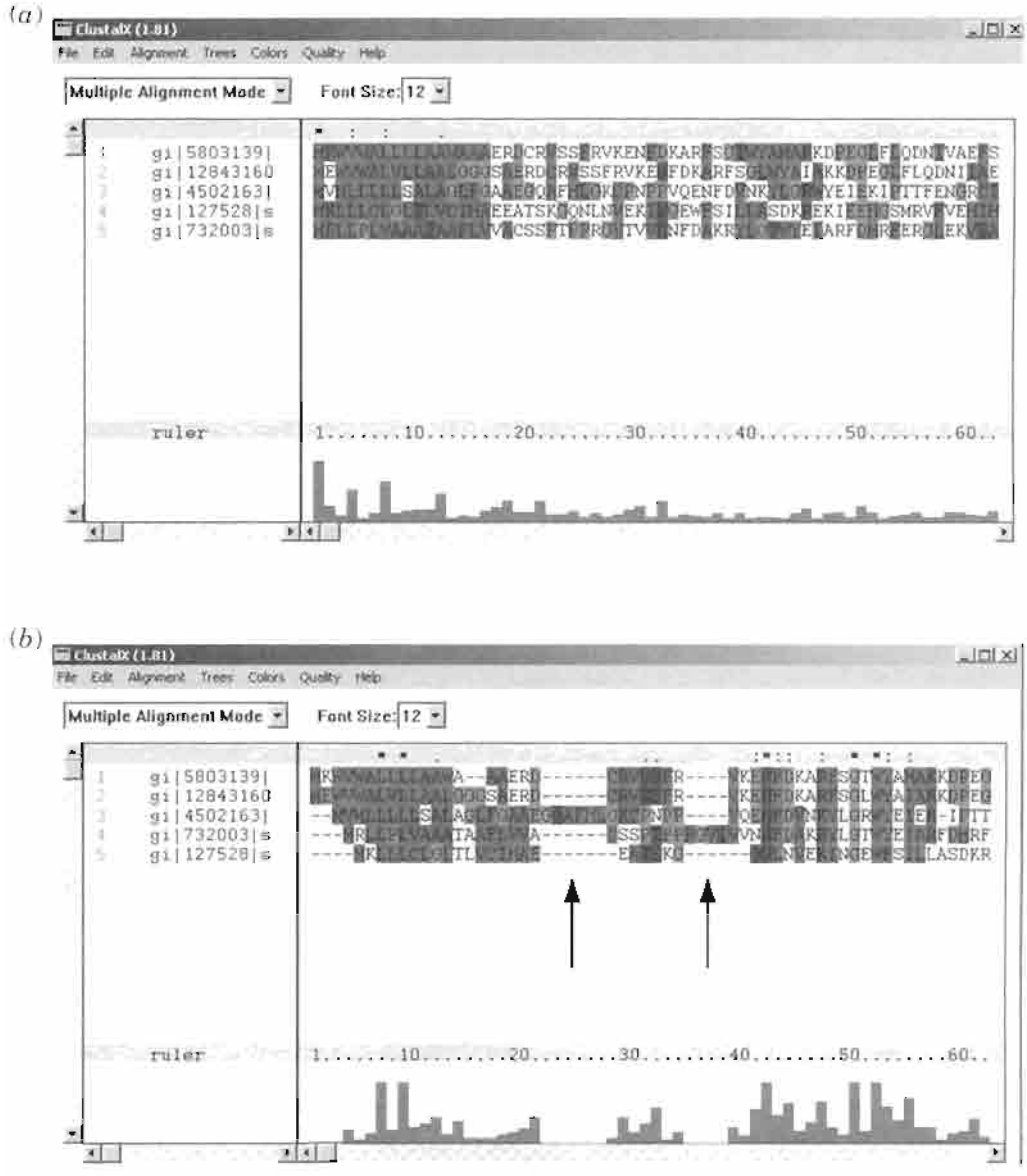


FIGURE 10.24. *The ClustalX program creates multiple sequence alignments. Five lipocalin protein sequences were obtained from the NCBI Entrez site in the FASTA format, copied into a word document, and saved as a text file (ASCII format) (see Fig. 10.1 for proteins). This file was then input into ClustalX with the "load sequences" command. (a) The sequences are initially entered into ClustalX without gap insertions. (b) Using the "Do Complete Alignment" command under the Alignment pull-down menu, a multiple sequence alignment is produced. This can be edited manually. Asterisks above the aligned columns indicate positions of 100% amino acid identity. The histogram at the bottom of the display box indicates the relative amount of sequence conservation across the alignment. ClustalX alignments can be saved for phylogenetic analysis using other software such as PAUP (Chapter 11).*

should evaluate whether those two short, neighboring insertion sequences should be combined and overlaid to provide a more parsimonious alignment; in this case, they share too little sequence identity to justify such a change.

An introductory tutorial for using ClustalX in conjunction with phylogeny software has been written by Hall (2001).

PileUp (GCG)

The GCG package (see Appendix) offers PileUp, a multiple sequence alignment program. Enter the sequences of these five related lipocalins into the sequence editor of GCG:

1. Rat odorant-binding protein (P08937), a lipocalin that is secreted from the lateral nasal gland into mucus and tears.
2. Bovine odorant-binding protein (P07435), an ortholog of rat OBP that has similar odorant-binding properties and a similar three-dimensional structure. Surprisingly, the two proteins share only 30% sequence identity.
3. Aphrodisin, an aphrodisiac from the vaginal secretions of the female hamster *Cricetus cricetus* (I48075). This protein induces males to copulate.
4. Probasin, a lipocalin of unknown function secreted from epithelial cells of the prostate (NP_061998).
5. Bovine dander major allergen (Q28133) (Mantyjarvi et al., 1996).

Sequences such as these can be entered into GCG through the sequence editor (Seqed). The sequences can be selected as a group by creating a file (Fig. 10.25*a*). The group of proteins is then multiply aligned with the PileUp program using the command "pileup @lipocalins" (Fig. 10.25*b*). The multiple sequence alignment is shown in Figure 10.26*a*. The program allows the gap creation penalty and gap extension penalty to be adjusted; when these parameters are lowered, more gaps occur in the alignment (Fig. 10.26*b*). As a biologist, you need to evaluate the alignments to assess when it is reasonable to add or remove gaps. In this particular example, the alignment in Fig. 10.26*a* is probably superior because the alignment of most critical cysteines as well as the GXW motif is conserved.

Many options are possible (and can be viewed with the command "genhelp pileup"). For example, the order of the aligned proteins can be constrained in any way with the command "nosort." As a default, the algorithm will otherwise align the two most closely related sequences at the top (in this case, rat OBP and hamster aphrodisin). PileUp can align up to 500 sequences with lengths up to 7000 amino acids or nucleotides.

Whenever possible, explore the 3D structures of the proteins you are studying (see Chapter 9) to help evaluate the multiple sequence alignment.

Other Multiple Sequence Alignment Software

Many other multiple sequence alignment programs have been written. An overview is provided in Figure 10.27 (based upon Thompson et al., 1999a). The programs include Dialign (Morgenstern, 1999), hmmt, Mlpima (Smith and Smith, 1992), Multal, Multalign, PRRP, SAGA (Notredame and Higgins, 1996), and Sbpima (Smith and Smith, 1992). They can be categorized according to their use of global versus local alignment and progressive (Feng-Doolittle) or iterative algorithms in which an initial alignment is subsequently refined.

These various programs are not redundant because the algorithms generate different alignments starting with the same input sequences. Also, they offer a variety of features that help optimize your alignment. The Dialign and Multalin programs, for example, provide measures of confidence in the robustness of the alignment (Figs. 10.28 and 10.29).

(*a*)

```
> cat>lipocalins

rnobp.pep
btobp.pep
ccaphr.pep
rnprob.pep
btdma.pep
```

(*b*)

```
> pileup @lipocalins

PileUp creates a multiple sequence alignment from a group of related
sequences using progressive, pairwise alignments.  It can also plot a
tree showing the clustering relationships used to create the alignment.


   1       rnobp.pep   172 aa
   2       btobp.pep   159 aa
   3      ccaphr.pep   167 aa
   4      rnprob.pep   177 aa
   5       btdma.pep   172 aa

 What is the gap creation penalty (* 8 *) ?

 What is the gap extension penalty (* 2 *) ?

 This program can display the clustering relationships graphically.
 Do you want to:

     A) Plot to a FIGURE file called "pileup.figure"
     B) Plot graphics on HP7550 attached to /dev/tty15
     C) Suppress the plot

 Please choose one (* A *):

 The minimum density for a one-page plot is 4.7 sequences/100 platen units.
 What density do you want (* 4.7 *) ?

 What should I call the output file name (* lipocalins.msf *) ?

 Determining pairwise similarity scores...

   1   x     2       1.32
   1   x     3       1.89
   1   x     4       1.44
   1   x     5       1.48
   2   x     3       1.10
   2   x     4       1.16
   2   x     5       1.33
   3   x     4       1.48
   3   x     5       1.54
   4   x     5       1.48

 Aligning...

   1      ........-.
   2      ........-.
          ........-.
   3      ........-.
          ........-.
   4      .......-.

 FIGURE instructions are now being written into pileup.figure.

       Total sequences:          5
      Alignment length:        179
              CPU time:      00.27

 Output file:/nfs/groucho/user/pevsner/pevsner/bioinf.dir/lipocalins.msf
```

FIGURE 10.25. *The PileUp program of GCG performs multiple sequence alignments of protein or DNA. Sequences can be copied and pasted into the sequence editor Seqed. (a) Sequences to be aligned are gathered in a file (called lipocalins). (b) The program is run by typing "pileup @lipocalins." PileUp gives you the option to change the gap creation and gap extension penalties; by lowering them (e.g., to values of 6 and 1), more gaps are typically introduced, but divergent sequences are better aligned. The output options include an alignment (the default option) or the guide tree used to make the alignment. As with other progressive alignment programs, PileUp creates pairwise similarity scores. The output is in the multiple sequence format (MSF) (Fig. 10.26). Other programs such as Pfam (this chapter) or PAUP (Chapter 11) use MSF as input or output options.*

Assessment of Alternative Multiple Sequence Alignment Algorithms

Several groups have systematically compared various multiple sequence alignment algorithms to assess their accuracy and performance properties (Morgenstern et al., 1996; McClure et al., 1994; Thompson et al., 1999a; Gotoh, 1996; Briffeuil et al., 1998). The approach is to obtain (or create) a database of reference alignments derived from protein sequences with known structures. Thus one can study proteins which are by definition structurally homologous. This allows an assessment of how successfully assorted multiple sequence alignment algorithms are able to detect

(*a*)

```
> cat lipocalins.msf
!!AA_MULTIPLE_ALIGNMENT 1.0
PileUp of: @lipocalins

 Symbol comparison table: GenRunData:blosum62.cmp  CompCheck: 6430

                  GapWeight: 8
            GapLengthWeight: 2

 lipocalins.msf  MSF: 179  Type: P  September 2, 2001 17:46  Check: 6778

 Name: rnobp           Len:   179  Check: 2583  Weight:  1.00
 Name: ccaphr          Len:   179  Check: 6725  Weight:  1.00
 Name: btdma           Len:   179  Check: 4045  Weight:  1.00
 Name: rnprob          Len:   179  Check: 3848  Weight:  1.00
 Name: btobp           Len:   179  Check: 9577  Weight:  1.00

//

        1                                                    50
 rnobp  MVKFLLIVLA LGVSC.AHHE NLDISPSEVN GDWRTLYIVA DNVEKVAEGG
ccaphr  MVKILL..LA LVFSL.AHAQ DF...AELQ GKWYTIVIAA DNLEKIEEGG
 btdma  ~MKAVFLTLL FGLVCTAQET PAEIDPSKIP GEWRIIYAAA DNKDKIVEGG
rnprob  ~MRVILLLLT LDVLGVSSMM TDKNLKKKIE GNWRTVYLAA SSVEKINEGS
 btobp  ~~~~~~~~~~ ~~~~~~~AQEE EAEQNLSELS GPWRTVYIGS TNPEKIQENG

        51                                                  100
 rnobp  SLRAYFQHME CGDECQELKI IFNVKLDSEC QTHTVVGQKH EDGRYTTDYS
ccaphr  PLRFYFRHID CYKNCSEMEI TFYVITNNQC SKTTVIGYLK GNGTYQTQFE
 btdma  PLRNYYRRIE CINDCESLSI TFYLKDQGTC LLLTEVAKRQ EGYVYVLEFY
rnprob  PLRTYFRRIE CGKRCNRINL YFYIKKGAKC QQFKIVGRRS QD.VYYAKYE
 btobp  PFRTYFRELV FDDEKGTVDF YFSVKRDGKW KNVHVKATKQ DDGTYVADYE

        101                                                 150
 rnobp  GRNYFHVLKK TDDIIFFHNV NVDESGRR.Q CDLVAGKRED LNKAQKQELR
ccaphr  GNNIFQPLYI TSDKIFFTNK NMDRAGQETN MIVVAGKGNA LTPEENEILV
 btdma  GTNTLEVIHV SENMLVTYVE NYD.GERITK MTEGLAKGTS FTPEELEKYQ
rnprob  GSTAFMLKTV NEKILLFDYF NRNRRNDVTR VAGVLAKGRQ LTKDEMTEYM
 btobp  GQNVFKIVSL SRTHLVAHNI NVDKHGQTTE LTELFVKLN. VEDEDLEKFW

        151                                 179
 rnobp  KLAEEYNIPN ENTQHLVPTD TCNQ~~~~~~
ccaphr  QFAHEKKIPV ENILNILATD TCPE~~~~~~
 btdma  QLNSERGVPN ENIENLIKTD NCPP~~~~~~
rnprob  NFVEEMGIED ENVQRVMDTD TCPNKIRIR
 btobp  KLTEDKGIDK KNVVNFLENE DHPHPE~~~~
```

(*b*)

```
        1                                                    50
 rnobp  M.VKFLLIVL .ALGV.SCAH HENLDISPSE VNGDWRTLYI VADNVEKVAE
ccaphr  M.VKILL..L .ALVF.SLAH AQ..DF..AE LQGKWYTIVI AADNLEKIEE
rnprob  MRVILLLLTL DVLGVSSMMT DKNL...KKK IEGNWRTVYL AASSVEKINE
 btdma  MKAVFL..TL .LFGLVCTAQ ETPAEIDPSK IPGEWRIIYA AADNKDKIVE
 btobp  ~~~~~~~~~~ ~~~~~~~~~AQ EEEAEQNLSE LSGPWRTVYI GSTNPEKIQE

        51                                                  100
 rnobp  GGSLRAYFQH MECGDECQEL KIIFNVKLDS ECQTHTVVGQ KHEDGRYTTD
ccaphr  GGPLRFYFRH IDCYKNCSEM EITFYVITNN QCSKTTVIGY LKGNGTYQTQ
rnprob  GSPLRTYFRR IECGKRCNRI NLYFYIKKGA KCQQFKIVGR RSQD.VYYAK
 btdma  GGPLRNYYRR IECINDCESL SITFYLKDQG TCLLLTEVAK RQEGYVYVLE
 btobp  NGPFRTYFRE LVFDDEKGTV DFYFSVKRDG KWKNVHVKAT KQDDGTYVAD

        101                                                 150
 rnobp  YSGRNYFHVL KKTDDII..F .FHNVNVDES GRRQCDL..V AG...KREDL
ccaphr  FEGNNIFQPL YITSDKI..F .FTNKNMDRA G.QETNMIVV AG...KGNAL
rnprob  YEGSTAF.ML KTVNEKILLF DYFNRN.... ...RRNDVTRV AGVLAKGRQL
 btdma  FYGTNTLEVI HVSENMLVTY V...ENYD.. GERITKMTE. .G.LAKGTSF
 btobp  YEGQNVFKIV SLSRTHLVA. ..HNINVDKH G.QTTELTEL ...FVKLN.V

        151                                 188
 rnobp  NKAQKQELRK LAEEYNIPNE NTQHLVPTDT CNQ~~~~~~
ccaphr  TPEENEILVQ FAHEKKIPVE NILNILATDT CPE~~~~~~
rnprob  TKDEMTEYMN FVEEMGIEDE NVQRVMDTDT CPNKIRIR
 btdma  TPEELEKYQQ LNSERGVPNE NIENLIKTDN CPP~~~~~~
 btobp  EDEDLEKFWK LTEDKGIDKK NVVNFLENED HPHPE~~~~
```

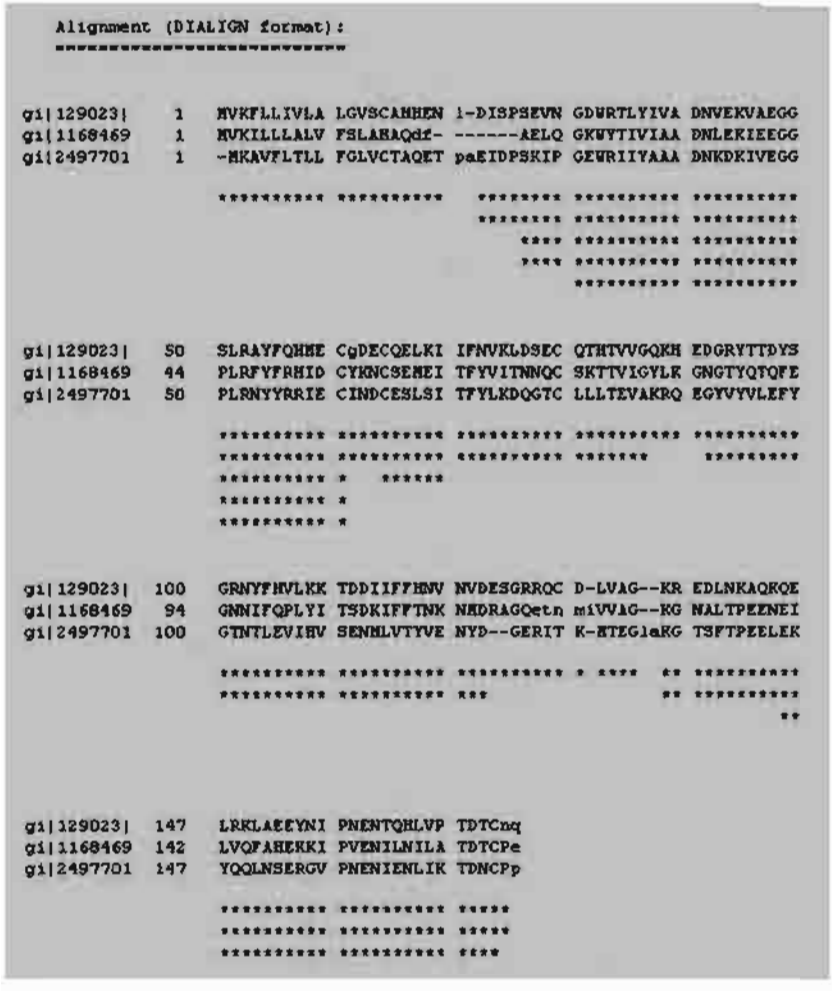
FIGURE 10.26. *Multiple sequence alignment of five lipocalins using the PileUp program (see Appendix). (a) The default result. There are several internal gaps (periods) and terminal gaps (tildes) in the alignment. (b) The result of repeating the alignment with a lower gap creation penalty (4) and gap extension penalty (1). Note that many additional gaps appear in the alignment. Regions of ambiguous alignment are usually deleted before performing phylogenetic analyses (Chapter 11).*

distant relationships among proteins. Thompson and colleagues (1999b) created BAliBASE as a reference alignment resource with over 1000 sequences in 142 alignments (▸http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE/index.html). The HIV-1 protease alignment is shown in Figure 10.30.

There are several conclusions that can be drawn from these comparative studies:

- As the group of sequences being multiply aligned begins to share less amino acid identity, the accuracy of the alignments decreases (Briffeuil et al., 1998).

FIGURE 10.27. *Overview of multiple sequence alignment algorithms as summarized by Thompson et al. (1999a). Progressive alignment algorithms use the method of Feng and Doolittle (1987). Iterative strategies refine an initial alignment. UPGMA and neighbor joining refer to types of guide trees (Chapter 11). Used with permission.*



```
   Alignment (DIALIGN format):
   ---------------------------


gi|129023|    1   MVKFLLIVLA LGVSCAHHEN l-DISPSEVN GDWRTLYIVA DNVEKVAEGG
gi|1168469    1   MVKILLLALV FSLAHAQdf- ------AELQ GKWYTIVIAA DNLEKIEEGG
gi|2497701    1   -MKAVFLTLL FGLVCTAQET paEIDPSKIP GEWRIIYAAA DNKDKIVEGG

                  ********** **********   ********* ********** **********
                                         ********* ********** **********
                                             **** ********** **********
                                             **** ********** **********
                                                  ********** **********


gi|129023|   50   SLRAYFQHEE CqDECQELKI IFNVKLDSEC QTHTVVGQKH EDGRYTTDYS
gi|1168469   44   PLRFYFRHID CYKNCSEMEI TFYVITNNQC SKTTVIGYLK GNGTYQTQFE
gi|2497701   50   PLRNYYRRIE CINDCESLSI TFYLKDQGTC LLLTEVAKRQ EGYVYVLEFY

                  ********** ********** ********** ********** **********
                  ********** ********** ********** ********     **********
                  ********** *    ******
                  ********** *
                  ********** *


gi|129023|  100   GRNYFHVLKK TDDIIFFHNV NVDESGRRQC D-LVAG--KR EDLNKAQKQE
gi|1168469   94   GNNIFQPLYI TSDKIFFTNK NMDRAGQetn miVVAG--KG NALTPEENEI
gi|2497701  100   GTNTLEVIHV SENHLVTYVE NYD--GERIT K-ETEGlaKG TSFTPEELEK

                  ********** ********** ********** * ****  ** **********
                  ********** ********** ***                ** **********
                                                                    **


gi|129023|  147   LRKLAEEYNI PNENTQHLVP TDTCnq
gi|1168469  142   LVQFAHERKI PVENILNILA TDTCPe
gi|2497701  147   YQQLNSERGV PNENIENLIK TDNCPp

                  ********** ********** *****
                  ********** ********** *****
                  ********** ********** ****
```

FIGURE 10.28. *Multiple sequence alignment of three lipocalins using the Dialign program (http://bibiserv.techfak.uni-bielefeld.de/dialign/). The proteins were rat odorant-binding protein (P08937), hamster aphrodisin (P09465), and bovine dander major allergen (Q28133). The program uses uppercase letters for aligned regions. Five asterisks indicate regions of maximal conservation. Identification of such regions can be useful for phylogenetic analyses.*

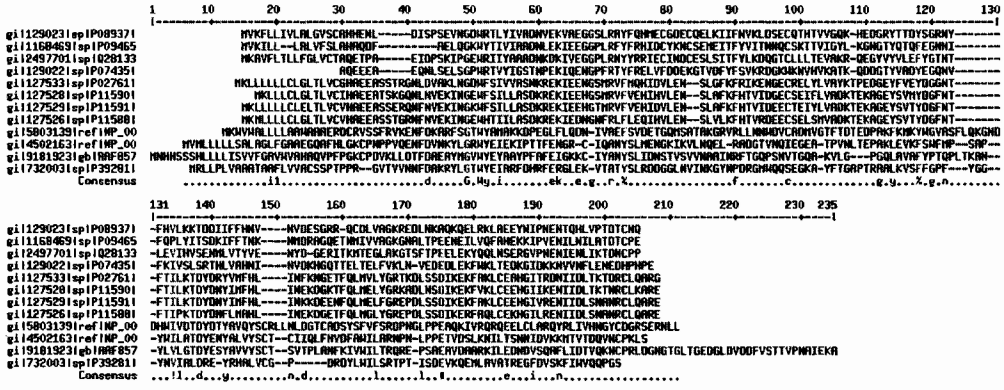


FIGURE 10.29. *The Multalin program of Florence Corpte performs multiple sequence alignments (▸http://prodes.toulouse.inra.fr/multalin/multalin.html). Aligned columns are color-coded red (high consensus), blue (low consensus), and black (neutral). The program offers a wide variety of input and output options such as assorted scoring matrices, gap penalty parameters, and color schemes.*

```
Name                    hiv-1 protease

Number of sequences              4
Alignment Length               106
Longest Sequence               104
Shortest Sequence               98
Average Percent Identity        49
Maximum Percent Identity        86
Minimum Percent Identity        35

Sequence Name    SWISSPROT Accession
1fmb                 P32542
7upjB                P03366
pol_sivcz            P17283
POL_SIVMK            P05897

Family   1fmb 7upjB pol_sivcz POL_SIVMK


1fmb          1  vTYNLEKRPTTIVLINDTPLNVLLDTGADTSVLTTahynr1kyrgrk.YQ
7upjB         1  pQFSLWKRPVVTAYIEGQPVEVLLDTGADDSIVAG.....iel.gnn.YS
pol_sivcz     1  pQITLWQRPLIPVKVEGQLCEALLDTGADDTVIER.....iqlqg1..WK
POL_SIVMK     1  pQFSLWRRPVVTAHIEGQPVEVLLDTGADDSIVTG.....iel.gph.YT

1fmb         50  GTGIGGVGGNVETFS.TPVTIKKKGRHIKTRMLVADIPVTILGRDILQDL
7upjB        44  PKIVGGIGGFINTLEYKNVEIEVLNKKVRATIMTGDTPINIFGRNILTAL
pol_sivcz    44  PKMIGGIGGFIKVKQFDNVHIEIEGRKVVGTVLVGPTPVNIIGRNILTQL
POL_SIVMK    44  PKIVGGIGGFINTKEYKNVEIEVLGKRIKRTIMTGDTPINIFGRNLLTAL

1fmb         99  GAKLV1
7upjB        94  GMSLN1
pol_sivcz    94  GCTLV.
POL_SIVMK    94  GMSLN1

Key

alpha helix          RED
beta strand          GREEN
core blocks          UNDERSCORE
```

You can also look at the alignment in *RSF format*, or *MSF format* with a *Feature Table*

FIGURE 10.30. *Example of a multiple sequence alignment from BAliBASE (▸http://www-igbmc.u-strasbg.fr/BioInfo/BAliBASE/index.html). This database contains reference alignments of varying degrees of amino acid identity. This is the alignment for HIV-1 protease, a group of sequences that are classified as short and >35% identity. There are 142 groups that are useful to evaluate the performance of many multiple sequence alignment algorithms.*

For groups of sequences that share less than 25% identity, the problem becomes especially severe. Thompson et al. (1999a) found that the best programs (PRRP, ClustalX, and SAGA) aligned about 60–70% of the amino acid residues for groups of proteins with <25% identity. For multiple sequence alignments of proteins sharing more identity (20% up to 40%), they found that on average 80% of the residues were aligned properly (Thompson et al., 1999a).

- Orphan sequences are proteins that are highly divergent members of a family. If we examined a multiple sequence alignment of RBP from 10 species, then added the OBP to that multiple sequence alignment, OBP would be considered an orphan. Orphans might be expected to disrupt the organization of a multiple sequence alignment, and yet they do not. Global alignment algorithms outperform local alignment methods for the introduction of orphans to an alignment (Thompson et al., 1999a).
- Separate multiple sequence alignments can be combined, such as a group of closely related RBPs and a group of closely related OBPs. Iterative algorithms such as PRRP and SAGA performed this task better than progressive alignment methods such as ClustalX (Thompson et al., 1999a). Considering all the comparative analyses, iterative strategies are not clearly superior to progressive sequence alignment methods.
- Often, some proteins in a family contain large extensions at the amino and/or carboxy termini. Local alignment programs dramatically outperformed global alignment programs at this task. One exception was PileUp, which uses a global algorithm but nonetheless performed well (Thompson et al., 1999a). In general, for most multiple sequence alignment applications, global alignments are superior.
- The effects of using different amino acid substitution matrices in multiple sequence alignment are relatively minor (Gotoh, 1996).

PERSPECTIVE

Multiple sequence alignment is the operation by which all the members of a protein family (or DNA family) may be grouped together. It is through multiple sequence alignment that we can identify and define the paralogs (family members within a species) and orthologs (family members between species) of any gene or gene product. Databases of multiply aligned protein families such as Pfam and InterPro are rapidly expanding in size and are increasingly important tools. These databases are often accompanied by careful expert annotation. A general trend is that databases offer the integration of many alignment resources. As genomic sequences rapidly accumulate, it becomes essential to use resources such as InterPro to functionally annotate predicted protein sequences.

There are many programs that generate multiple sequence alignments. These programs often produce dramatically different results depending on the nature of the sequences that are being aligned. It is reasonable to pick one program to use for your multiple sequence alignments, such as ClustalW, ClustalX, or PileUp. However, it is also reasonable to make the effort to adjust the parameters to optimize the alignments.

PITFALLS

There are several categories of errors associated with multiple sequence alignment. These include:

- False negatives: performing an alignment that omits true homologs.
- False positives: adding family members that are not authentic.
- Alignment errors: failing to match distant subgroups.
- Alignment errors: adding gaps improperly.

It is especially important to perform a proper multiple sequence alignment for molecular phylogeny studies. The alignment constitutes the raw data that go into making a tree (see Chapter 11).

For popular programs such as PileUp, the global alignment algorithm sometimes results in the total failure to align a distantly related sequence to other proteins. This problem can be avoided by adjusting the gap creation and/or gap extension penalties.

WEB RESOURCES

Multiple sequence alignments are used to define which proteins cluster into families. In this chapter, we divided the programs into those databases that contain prefabricated alignments (Table 10.1) and those algorithms that let a user generate a customized alignment (Table 10.3).

DISCUSSION QUESTIONS

[10-1] Feng and Doolittle introduced the "once a gap, always a gap" rule, saying that the two most closely related sequences that are initially aligned should be weighted most heavily in assigning gaps. Why was it necessary to introduce this rule?

[10-2] Could BLAST searches incorporate HMMs? How does PSI-BLAST differ from an HMM-based search in Pfam?

PROBLEMS

[10-1] X-linked adrenoleukodystrophy (X-ALD) is the most common inherited disease affecting peroxisomes (a subcellular organelle involved in lipid metabolism and other metabolic functions). The disease is caused by mutations in the ABCD1 gene on chromosome Xq28 encoding ALD protein (ALDP). In humans, there are thought to be four ALDP-related proteins on peroxisomes: ALDP (NP_000024; 745 amino acid residues), ALDR (NP_005155, 740 residues), PMP70 (NP_002849, 659 residues), and PMP70R (NP_005041, 606 residues). Two yeast ALDP-like proteins have also been identified, Pxa1p (NP_015178) and Pxa2p (NP_012733). These proteins are all part of a much larger family of ATP-binding cassette (ABC) transporters, including the cystic fibrosis transmembrane regulator (CFTR) and multidrug-resistant proteins (MDR).

Create a multiple sequence alignment of the human, mouse, and yeast ALDP family of proteins. Identify the presumed nucleotide binding site, GPNGCGKS. Is this motif perfectly conserved?

[10-2] The quagga was an African animal that is now extinct. It looked partly like a horse and partly like a zebra. In 1872, the last living quagga was photographed (Fig. 10.31). Mitochondrial DNA was obtained from a museum specimen of a quagga and sequenced. Perform a multiple sequence alignment of quagga (*Equus quagga boehmi*), horse (*Equus caballus*), and zebra (*Equus burchelli*) mitochondrial DNA. To which animal was the quagga more closely related?

[10-3] In order to determine which spirochetes invaded the gums of a patient with severe periodontitis, researchers

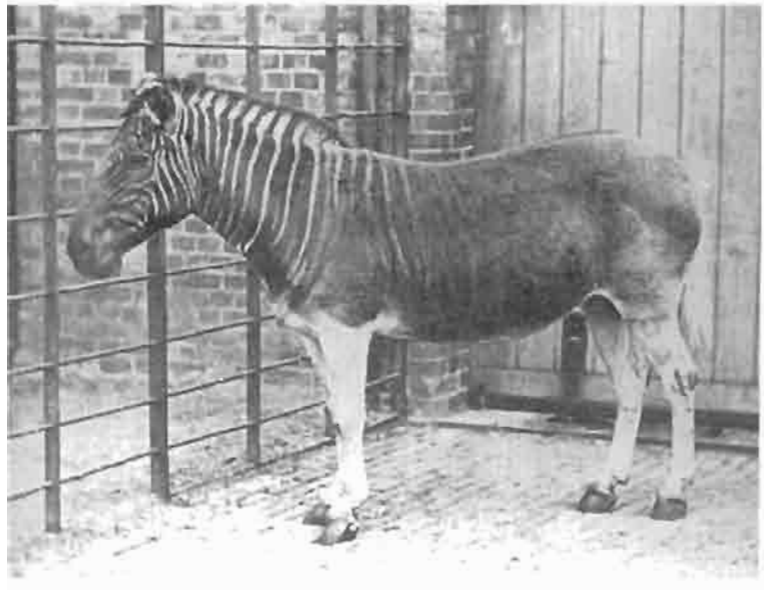


FIGURE 10.31. Photograph of a quagga from the London Zoo (1872). Was this extinct animal more closely related to the horse or the zebra? This photograph is available on the internet (▸www.oursci.org/magazine/200108/010819.htm;homepages.tesco.net/~zoechaos/quaggas.htm; ▸www.greenapple.com/~jorp/amzanim/quagga.htm;www.foek.hu/dodo/emlos/eqqu.htm).

cloned DNA and RNA samples from the subgingival plaque (Choi et al., 1994). Examine the results in PopSet at the NCBI Entrez site. Identify a portion of the multiple sequence alignment that you can improve manually.

[10-4] Study a protein family in Pfam beginning with "text search."
(a) As an example, look at the ABC transporters again.
(b) Go to the list of the "top 20" Pfam families (▸http://▸Pfam.wustl.edu/browse.shtml).
Select one of these, and explore it. Try JalView, and look at the family by principal components analysis and by making a tree.

[10-5] You have identified the following accession numbers for lipocalins:

Human retinol binding protein	NP_006735
Human apolipoprotein D	NP_001638
Rat odorant-binding protein	P08937

Paste these sequences in FASTA format into ClustalW (▸http://www.ebi.ac.uk/clustalw/). Create a multiple sequence alignment.

SELF-TEST QUIZ

[10-1] Why doesn't ClustalW (a program that employs the Feng and Doolittle progressive sequence alignment algorithm) report expect values?
(a) ClustalW *does* report expect values.
(b) ClustalW uses global alignments for which *E* value statistics are not available.
(c) ClustalW uses local alignments for which E value statistics are not available.
(d) *E* value statistics are not relevant to multiple sequence alignments.

[10-2] Which of the following programs does NOT generate a multiple sequence alignment?
(a) PSI-BLAST
(b) ClustalW
(c) PileUp
(d) PHYLIP

[10-3] The "once a gap, always a gap" rule for the Feng–Doolittle method:
(a) Assures that gaps will not be filled in inappropriately with inserted sequences
(b) Assures that sequences that diverged early in evolution will be given priority in establishing the order in which a multiple sequence alignment is constructed
(c) Assures that gaps occurring between sequences that are most closely related in a multiple sequence alignment will be preserved
(d) Assures that gaps occurring between sequences that are distantly related will be maintained in the multiple sequence alignment

[10-4] Which of the following statements best illustrates the theory behind the hidden Markov model (HMM)?
(a) It relies on first creating a phylogenetic tree.
(b) It calculates the probability of an amino acid occurrence at each position.
(c) It calculates a multiple sequence alignment based on scores from randomly generated sequences.
(d) It only aligns sequences that belong to an already described protein family.

[10-5] Which of the following programs compares a protein query to a set of many position-specific scoring matrices?

(a) PHI-BLAST
(b) RPS-BLAST
(c) PROSITE
(d) ProDom

[10-6] Which of the following is not a database consisting primarily of hidden Markov models?
(a) Pfam
(b) PRINTS
(c) SMART
(d) TIGRFAMs

[10-7] A main advantage of searching the InterPro database is that:
(a) Almost 25% of SwissProt and TrEMBL protein sequences have one or more hits in InterPro.
(b) The InterPro member databases all use HMMs to define protein families.
(c) During sequence similarity searches, more than one domain may be present in a protein. InterPro is a protein signature database that is useful to characterize these domains.
(d) During sequence similarity searches, you may want to search a reverse-profile HMM against one domain of your query. InterPro is useful to accomplish this.

[10-8] If you perform a multiple sequence alignment of a group of proteins and include a distantly related protein (a divergent member called an "orphan"):
(a) The orphan is typically aligned with the group of proteins.
(b) The orphan is typically not aligned with the group of proteins.

[10-9] The main difference between Pfam-A and Pfam-B is that:
(a) Pfam-A is manually curated while Pfam-B is automatically curated.
(b) Pfam-A uses hidden Markov models while Pfam-B does not.
(c) Pfam-A provides full-length protein alignments while Pfam-B aligns protein fragments.
(d) Pfam-A incorporates data from SMART and PORSITE while Pfam-B does not.

SUGGESTED READING

Da-Fei Feng and Russell F. Doolittle's (1987) progressive alignment approach to multiple sequence alignment is an important paper. This work stresses the relationship between multiple sequence alignment and the evolutionary relationships of proteins. It is thus relevant to our treatment of phylogeny in Chapter 11. Doolittle (2000) also wrote a personal account of his interest in sequence analysis, phylogeny, and bioinformatics, including mention of the historical context in which he developed his alignment algorithm.

InterPro and its constituent databases (PRINTS, PROSITE, Pfam, ProDom, SMART, TIGRFAMs, SwissProt, and TrEMBL) are the subject of an issue of *Briefings in Bioinformatics* (volume 3, number 3, September 2002). All of the articles in this issue offer excellent insights into these databases and into the importance of InterPro as a resource to integrate vast amounts of data on protein families and domains.

A few research groups have systematically compared multiple sequence alignment algorithms, and reading any of these papers provides a deeper insight into the strengths and weaknesses of the algorithms (Briffeuil et al., 1998; Gotoh, 1996; McClure et al., 1994; Park et al., 1998; Thompson et al., 1999a). A study by Julie Thompson and colleagues (1999a) is particularly informative. A paper by Park et al. (1998) is particularly useful to explore the usefulness of PSI-BLAST (Chapter 5) relative to other, related algorithms.

A series of articles on multiple sequence alignment appears in *Methods in Enzymology* (volume 266, 1996) with articles by William Taylor, Da-Fei Feng and Russell Doolittle, Desmond Higgins, and others.

For HMMs, a number of excellent introductions were cited in this chapter, such as a brief review by Ewan Birney (2001).

REFERENCES

Apweiler, R., et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29,** 37–40 (2001).

Attwood, T. K., et al. PRINTS prepares for the new millennium. *Nucleic Acids Res.* **27,** 220–225 (1999).

Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. A. Hidden Markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. USA* **91,** 1059–1963 (1994).

Bateman, A., et al. The Pfam protein families database. *Nucleic Acids Res.* **28,** 263–266 (2000).

Bateman, A., et al. The Pfam protein families database. *Nucleic Acids Res.* **30,** 276–280 (2002).

Birney, E. Hidden Markov models in biological sequence analysis. *IBM J. Res. Dev.* **45,** 449–459 (2001).

Briffeuil, P., et al. Comparative analysis of seven multiple protein sequence alignment servers: Clues to enhance reliability of predictions. *Bioinformatics* **14,** 357–366 (1998).

Choi, B. K., Paster, B. J., Dewhirst, F. E., and Gobel, U. B. Diversity of cultivable and uncultivable oral spirochetes from a patient with severe destructive periodontitis. *Infect. Immun.* **62,** 1889–1895 (1994).

Chothia, C., and Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5,** 823–826 (1986).

Doolittle, R. F. On the trail of protein sequences. *Bioinformatics* **16,** 24–33 (2000).

Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14,** 755–763 (1998).

Feng, D. F., and Doolittle, R. F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25,** 351–360 (1987).

Feng, D. F., and Doolittle, R. F. Progressive alignment and phylogenetic tree construction of protein sequences. *Methods Enzymol.* **183,** 375–387 (1990).

Gotoh, O. Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments. *J. Mol. Biol.* **264,** 823–838 (1996).

Hall, B. G. *Phylogenetic Trees Made Easy. A How-To for Molecular Biologists*. Sinauer Associates, Sunderland, MA, 2001.

Henikoff, S., Henikoff, J. G., and Pietrokovski, S. Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **15,** 471–479 (1999).

Henikoff, S., Pietrokovski, S., and Henikoff, J. G. Superior performance in protein homology detection with the Blocks Database servers. *Nucleic Acids Res.* **26,** 309–312 (1998).

Hermankova, M., et al. HIV-1 drug resistance profiles in children and adults with viral load of <50 copies/ml receiving combination therapy. *JAMA* **286,** 196–207 (2001).

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27,** 215–219 (1999).

Krogh, A. An introduction to hidden Markov models for biological sequences. *In Computational Methods in Molecular Biology*. S. L. Salzberg, D. B. Searls, S. Kasif (eds.) Elsevier, New York, 1998, Chapter 4.

Krogh, A., Brown, M., Mian, I. S., Sjolander, K., and Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235,** 1501–1531 (1994).

Mantyjarvi, R., et al. Complementary DNA cloning of the predominant allergen of bovine dander: A new member in the lipocalin family. *J. Allergy. Clin. Immunol.* **97,** 1297–1303 (1996).

McClure, M. A., Vasi, T. K., and Fitch, W. M. Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol.* **11,** 571–592 (1994).

Morgenstern, B. DIALIGN 2: Improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15,** 211–218 (1999).

Morgenstern, B., Dress, A., and Werner, T. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. *Proc. Natl. Acad. Sci. USA* **93,** 12098–12103 (1996).

Mulder, N. J., et al. InterPro: An integrated documentation resource for protein families, domains, and functional sites. *Brief. Bioinform.* **3,** 225–235 (2002).

Mulder N. J., Apweiler R., Attwood T. K., Bairoch A., Barrell D., Bateman A., Binns D., Biswas M., Bradley P., Bork P., Bucher P., Copley R. R., Courcelle E., Das U., Durbin R., Falquet L., Fleischmann W., Griffiths-Jones S., Haft D., Harte N., Hulo N., Kahn D., Kanapin A., Krestyaninova M., Lopez R., Letunic I., Lonsdale D., Silventoinen V., Orchard S. E., Pagni M., Peyruc D., Ponting C. P., Selengut J. D., Servant F., Sigrist C. J., Vaughan R., Zdobnov E. M. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31,** 315–318 (2003).

Needleman, S. B., and Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48,** 443–453 (1970).

Notredame, C. and Higgins, D. G. SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res.* **24,** 1515–1524 (1996).

Park, J., et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284,** 1201–1210 (1998).

Parry-Smith, D. J., Payne, A. W., Michie, A. D., and Attwood, T. K. CINEMA—a novel colour INteractive editor for multiple alignments. *Gene* **221,** GC57–63 (1998).

Ponting, C. P., Schultz, J., Milpetz, F., and Bork, P. SMART: Identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res.* **27,** 229–232 (1999).

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci. USA* **95,** 5857–5864 (1998).

Shoop, E., Silverstein, K. A., Johnson, J. E., and Retzel, E. F. MetaFam: A unified classification of protein families. II. Schema and Query Capabilities. *Bioinformatics* **17,** 262–271 (2001).

Sigrist, C. J., et al. PROSITE: A documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* **3,** 265–274 (2002).

Silverstein, K. A., Shoop, E., Johnson, J. E., and Retzel, E. F. MetaFam: A unified classification of protein families. I. Overview and statistics. *Bioinformatics* **17,** 249–261 (2001a).

Silverstein, K. A., et al. The MetaFam Server: A comprehensive protein family resource. *Nucleic Acids Res.* **29,** 49–51 (2001b).

Smith, R. F., and Smith, T. F. Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Eng.* **5,** 35–41 (1992).

Thompson, J. D., Higgins, D. G., and Gibson, T. J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22,** 4673–4680 (1994).

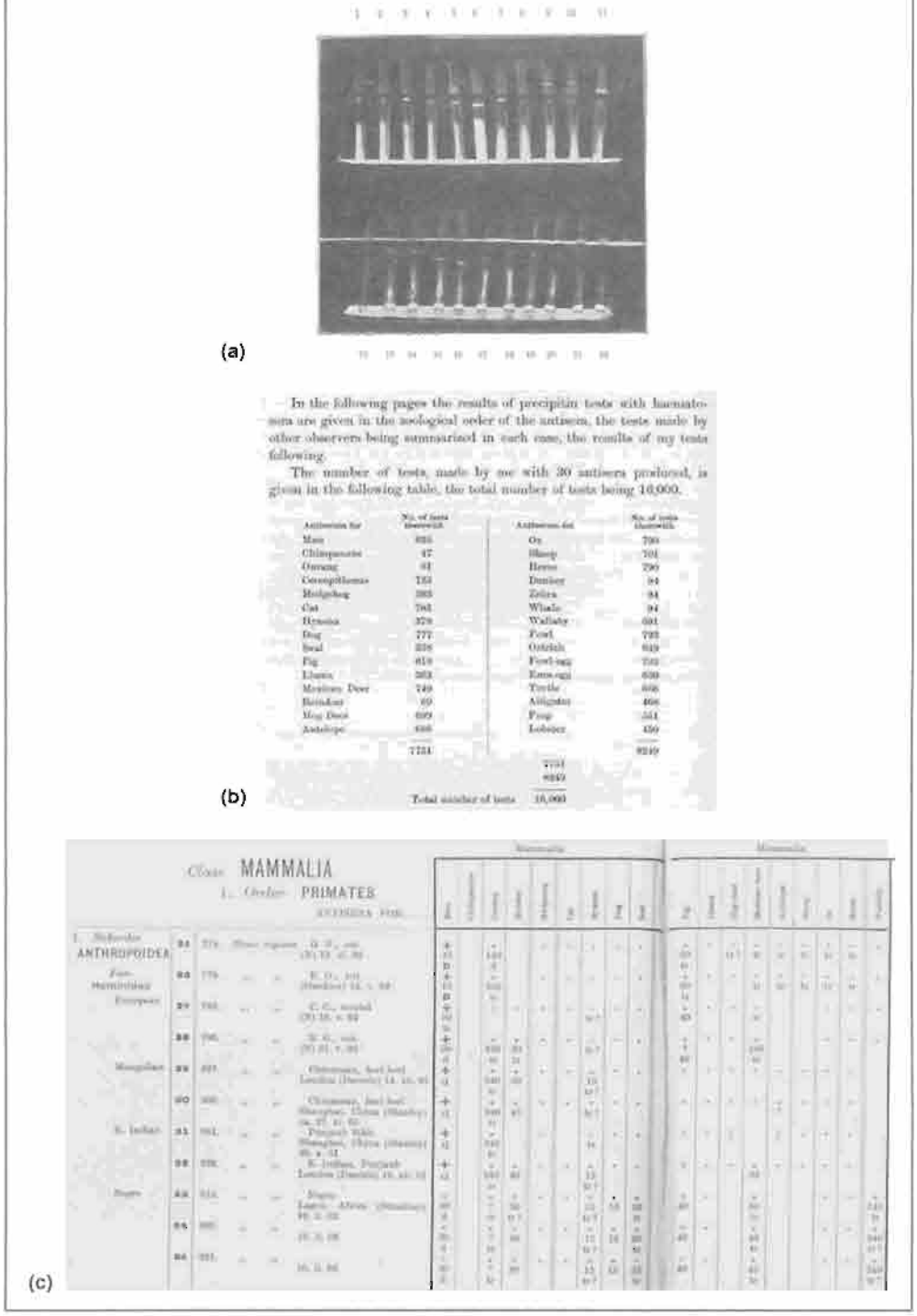
Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25,** 4876–4882 (1997).

Thompson, J. D., Plewniak, F., and Poch, O. A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.* **27,** 2682–2690 (1999a).

Thompson, J. D., Plewniak, F., and Poch, O. BAliBASE: A benchmark alignment database for the evaluation of multiple alignment programs. *Bioinformatics* **15,** 87–88 (1999b).

Tuppy, H. Über die Artspezifität der Proteinstruktur. In A. Neuberger (ed.) Symposium on Protein Structure. John Wiley & Sons, New York, 1958, pp. 66–76.

Wu, C. H., Xiao, C., Hou, Z., Huang, H., and Barker, W. C. iProClass: An integrated, comprehensive and annotated protein classification database. *Nucleic Acids Res.* **29,** 52–54 (2001).



(a)

In the following pages the results of precipitin tests with haematosera are given in the zoological order of the antisera, the tests made by other observers being summarized in each case, the results of my tests following.

The number of tests made by me with 30 antisera produced, is given in the following table, the total number of tests being 16,000.

Antiserum for	No. of tests	Antiserum for	No. of tests
Man	895	Ox	700
Chimpanzee	47	Sheep	701
Orang	41	Horse	700
Cercopithecus	733	Donkey	94
Hedgehog	385	Zebra	94
Cat	761	Whale	94
Hyæna	378	Wallaby	691
Dog	777	Fowl	723
Seal	398	Ostrich	619
Pig	619	Fowl-egg	702
Llama	383	Emu-egg	630
Mouse-Deer	749	Turtle	606
Reindeer	89	Alligator	466
Hog Deer	699	Frog	551
Antelope	466	Lobster	450
	7751		9249
			7751
			16,000
	Total number of tests		

(b)

Class MAMMALIA

1. Order PRIMATES

(c)

For the first half of the 20th century, the only phylogenetic analyses based on molecular data were the remarkable precipitin tests pioneered by George Nuttall and colleagues. Antisera were incubated with serum samples from a variety of species, and the time required for a precipitation reaction was recorded as well as the strength of the reaction. (a) *Sample test tubes in which the reactions were conducted (Nuttall, 1904, plate I).* (b) *Excerpt from Nuttall (1904, p. 160) describing the 16,000 tests he performed.* (c) *Portion of the 92-page data summary of Nuttall (1904, p. 222–223). The 900 rows (of which 11 are shown here) represent blood samples that were tested, and the columns correspond to antisera obtained from 30 organisms (of which 18 are shown here). The values represent the time (in minutes) required for a reaction. The symbols indicate the degree of reaction (+ being greatest, and · indicating no reaction). The letter D indicates the presence of deposits in the test tube. Nuttall used these data to infer the phylogenetic relationships of assorted mammals, birds, reptiles, amphibians, and crustaceans. In the 1950s and 1960s, amino acid sequence comparisons largely replaced immunological tests for phylogenetic analysis.*

11

Molecular Phylogeny and Evolution

Introduction to Molecular Evolution

Evolution is the theory that groups of organisms change over time so that descendants differ structurally and functionally from their ancestors. Evolution may also be defined as the biological process by which organisms inherit morphological and physiological features that define a species. In 1859 Charles Darwin published his landmark book, On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. "As many more individuals of each species are born than can possibly survive; and as, consequently, there is a frequently recurring struggle for existence, it follows that any being, if it vary however slightly in any manner profitable to itself, under the complex and sometimes varying conditions of life, will have a better chance of surviving, and thus be naturally selected. From the strong principle of inheritance, any selected variety will tend to propagate its new and modified form."

Evolution is a process of change. Heredity is generally conservative—offspring resemble their parents—and yet the structure and function of bodies changes over

We will explore the tree of life in Chapter 12.

You can read The Origin of Species by Charles Darwin on-line at ▸http://www.bbc.co.uk/education/darwin/origin/ or ▸http://www.literature.org/authors/darwin-charles/the-origin-of-species/.

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

the course of generations. There are three main mechanisms by which changes may occur (Simpson, 1952):

- Conditions of growth affect development. Environmental factors such as accidents and disease-causing infections are not hereditary in nature (although an individual's response to disease or environmental stimuli is genetically controlled to some extent, as discussed in Chapter 18.)
- The mechanism of sexual reproduction assures change from one generation to the next. Genes are "shuffled" in a unique combination when an offspring inherits chromosomes from two parents.
- Mutations can produce changes in genes and more generally in chromosomes.

At the molecular level, evolution is a process of mutation with selection. Molecular evolution is the study of changes in genes and proteins throughout different branches of the tree of life. This discipline also uses data from present-day organisms to reconstruct the evolutionary history of species.

Phylogeny is the inference of evolutionary relationships. Traditionally, phylogeny was assessed by comparing morphological features between organisms from a variety of species (Mayr, 1982). However, molecular sequence data can also be used for phylogenetic analysis. The evolutionary relationships that are inferred, which are usually depicted in the form of a tree, can provide hypotheses of past biological events.

The word phylogeny is derived from the Greek *phylon* ("race, class") and *geneia* ("origin"). Ernst Haeckel, whose tree of life is shown on the frontis to Chapter 12, coined the terms *phylogeny*, *phylum*, and *ecology*. He also wrote that "ontogenesis is a brief and rapid recapitulation of phylogenesis, determined by the physiological functions of heredity (generation) and adaptation (maintenance)" (Haeckel, 1900, p. 81). See also ▸http://www.ucmp.berkeley.edu/history/haeckel.html.

Historical Background

Tremendous progress was made in our understanding of molecular evolution through the study of insulin beginning in the 1950s. Insulin is a small protien that stimulates glucose uptake upon binding to an insulin receptor on pancreatic acinar cells. In 1953 Frederick Sanger and colleagues determined the primary amino acid sequence of insulin, the first time this feat had been accomplished for any protein. The mature, biologically active protein consists of two subunits, the A chain and B chain, that are covalently attached through intermolecular disulfide bridges. More recently, the structure of the human preproinsulin molecule was shown to consist of a signal peptide, the B chain, an intervening sequence called the C peptide, and the A chain (Fig. 11.1a). The C peptide is flanked by dibasic residues (arg-arg or lys-arg; see Fig. 11.1a,b) at which proteolytic cleavage occurs.

Frederick Sanger won the Nobel Prize in Chemistry (1958) "for his work on the structure of proteins, especially that of insulin" (▸http://www.nobel.se/chemistry/laureates/1958/). In 1980, he shared the Nobel Prize in Chemistry (with Paul Berg and Walter Gilbert) for his "contributions concerning the determination of base sequences in nucleic acids."

Sanger and others sequenced insulin proteins from five species (cow, sheep, pig, horse, and whale). It became clear immediately that the A chain and B chain residues are highly conserved. Furthermore, amino acid differences were restricted to three residues within a disulfide "loop" region of the A chain (Fig. 11.1b, shaded red). This suggested that amino acid substitutions occur nonrandomly, some changes affecting biological activity dramatically and other changes having negligible effects (Anfinsen, 1959). The differences within the disulfide loop are termed "neutral" changes (Jukes and Cantor, 1969, p. 86; Kimura, 1968). Later, when the biologically active A and B chain sequences were compared to the functionally less important C peptide, even more dramatic differences were seen. Kimura (1983) reported that the C peptide evolves at a rate of 2.4×10^{-9} per amino acid site per year, sixfold faster than the rate for the A and B chainss (0.4×10^{-9} per amino acid site per year). At the nucleotide level, the rate of evolution is similarly about sixfold faster for the DNA region encoding the C peptide (Li, 1997).

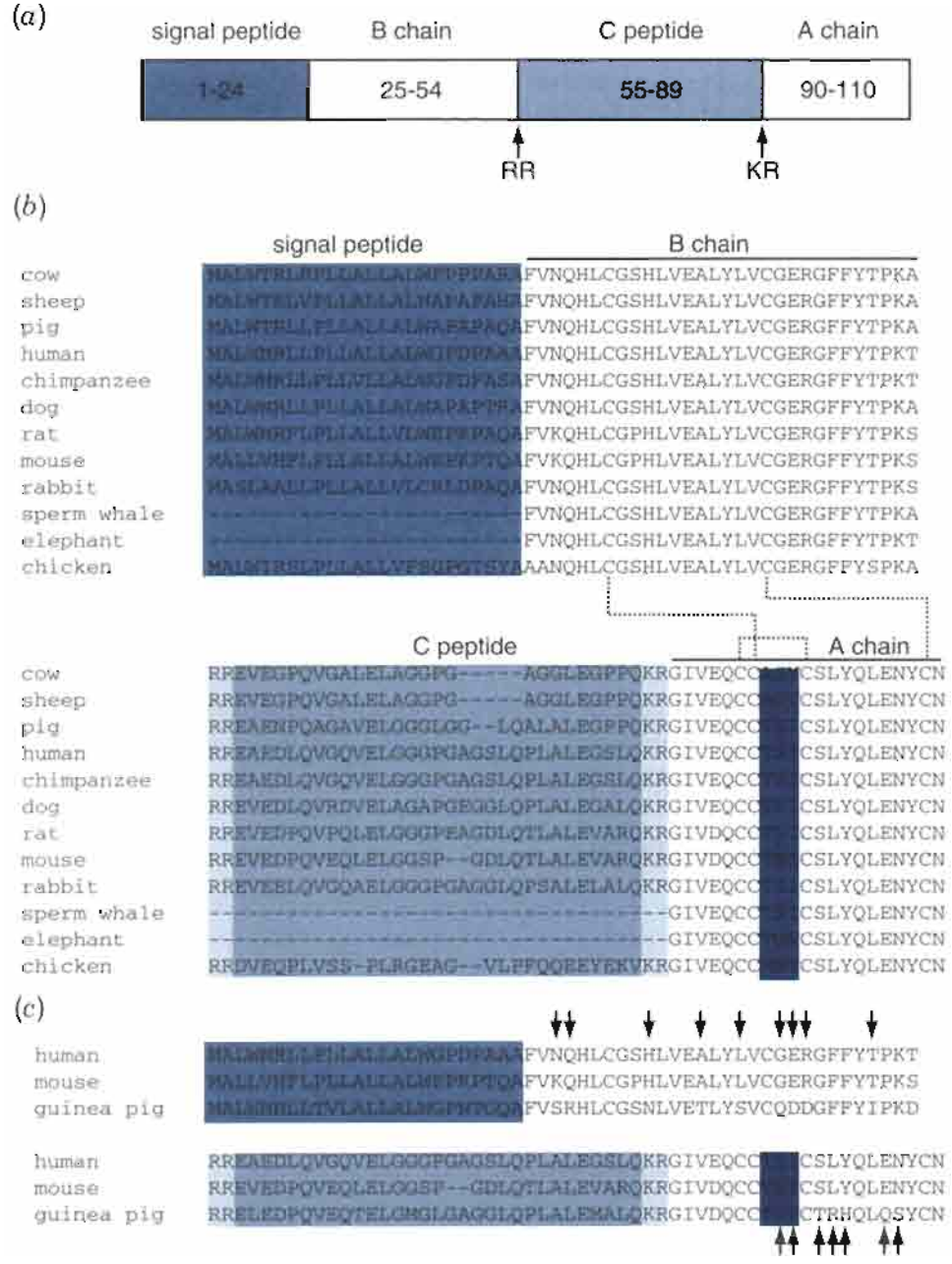


FIGURE 11.1. *Since the 1950s, studies of insulin have facilitated our understanding of molecular evolution. (a) The human insulin molecule consists of a signal peptide (required for intracellular transport; amino acid residues 1-24), the B chain, the C peptide, and the A chain. Dibasic residues (RR, KR) flank the C peptide and are the sites at which proteases cleave the protein. The A chain and B chain are then covalently linked through disulfide bridges, forming mature insulin. (b) Multiple sequence alignment of insulin from 12 species. Amino acid substitutions occur in nonrandom patterns. Note that within the A chain of insulin the amino acid residues are almost perfectly conserved between different species, except for three divergent columns of amino acids (A chain, colored region in a "disulfide loop"). However, the rate of nucleotide substitution is about sixfold higher in the region encoding the intervening C peptide than in the region encoding the B and A chains (Kimura, 1983), and gaps in the multiple sequence alignment are evident here. Disulfide bridges between cysteine residues are indicated by dashed lines. The accession numbers are NP_000198 (human), P30410 (chimpanzee), NP_062002 (rat), P01321 (dog), NP_032412 (mouse), P01311 (rabbit), P01315 (pig), P01332 (chicken), NP_776351 (cow), P01318 (sheep), INEL (elephant), and INWHP (sperm whale). (c) Guinea pig insulin (Cavia porcellus, accession P01329) evolves about sevenfold faster than insulin from other species. Human, mouse, and guinea pig insulins are aligned. Arrows indicate 16 amino acid positions at which the guinea pig sequence varies from that of human and/or mouse.*

Oxytocin CYIQNCPLG
Vasopressin CYFQNCPRG

FIGURE 11.2. *Oxytocin (LocusID 5020; NP_000906) and vasopressin (LocusID 551; NP_000481) differ at only two amino acid positions, yet they have vastly different biological functions. The comparison of these peptide sequences in the 1960s led to the appreciation of the importance of primary amino acid sequences in determining protein function.*

As insulin was sequenced from additional species, a surprising finding emerged. Insulin from guinea pig and a closely related species of the family *Caviidae* (the coypu) appeared to evolve seven times faster than insulin from other species. As shown in the alignment of Figure 11.1c, the guinea pig insulin sequence differs from human and mouse insulin at 16 different amino acid positions within the A and B chains. The explanation for this phenomenon (Jukes, 1979) is that guinea pig and coypu insulin do not bind two zinc ions, whereas insulin from all the other species do. There is presumably a strong functional constraint on most insulin molecules to maintain amino acid residues that are able to complex zinc. Guinea pig and coypu insulin have less constraint.

In the early 1950s, other laboratories sequenced vasopressin and oxytocin and found that peptides differing by only two amino acid residues have vastly different biological function (Fig. 11.2). And in 1960 Max Perutz and John Kendrew solved the structures of hemoglobin and myoglobin. These proteins, both of which serve as oxygen carriers, are homologous and share related structures. Thus it became clear by the 1960s that there are significant structural and functional consequences to variation in primary amino acid sequence.

Perutz and Kendrew won the 1962 Nobel Prize in Chemistry "for their studies of the structures of globular proteins." You can read about their accomplishments at ▸http://www.nobel.se/chemistry/laureates/1962/.

Molecular Clock Hypothesis

In the 1960s, primary amino acid sequence data were accumulated for abundant, soluble proteins such as hemoglobins, cytochrome *c*, and fibrinopeptides in a variety of species. Some proteins, such as cytochrome *c* from many organisms, were found to evolve very slowly, while other protein families accumulated many substitutions. Linus Pauling and others proposed the concept of a molecular clock (Zuckerkandl and Pauling, 1962; Margoliash, 1963). This hypothesis states that for every given gene (or protein), the rate of molecular evolution is approximately constant.

A study demonstrating the existence of a molecular clock was performed by Richard Dickerson in 1971 (Fig. 11.3). He analyzed three proteins for which a large amount of sequence data were available: cytochrome *c*, hemoglobin, and fibrinopeptides. For each, he plotted the relationship between the number of amino acid differences for a protein in two organisms versus the divergence time (in millions of years, MY) for the organisms. These divergence times were estimated from paleontology. The *y* axis of this plot consists of the corrected number of amino acid changes per 100 residues, *m*. The value of *m* is calculated

$$\frac{m}{100} = -\ln\left(1 - \frac{n}{100}\right) \tag{11.1}$$

This equation can be restated as

$$\frac{n}{100} = 1 - e^{-(m/100)} \tag{11.2}$$

where *m* is the total number of amino acid changes which have occurred in a 100-amino-acid segment of a protein and *n* is the observed number of amino acid

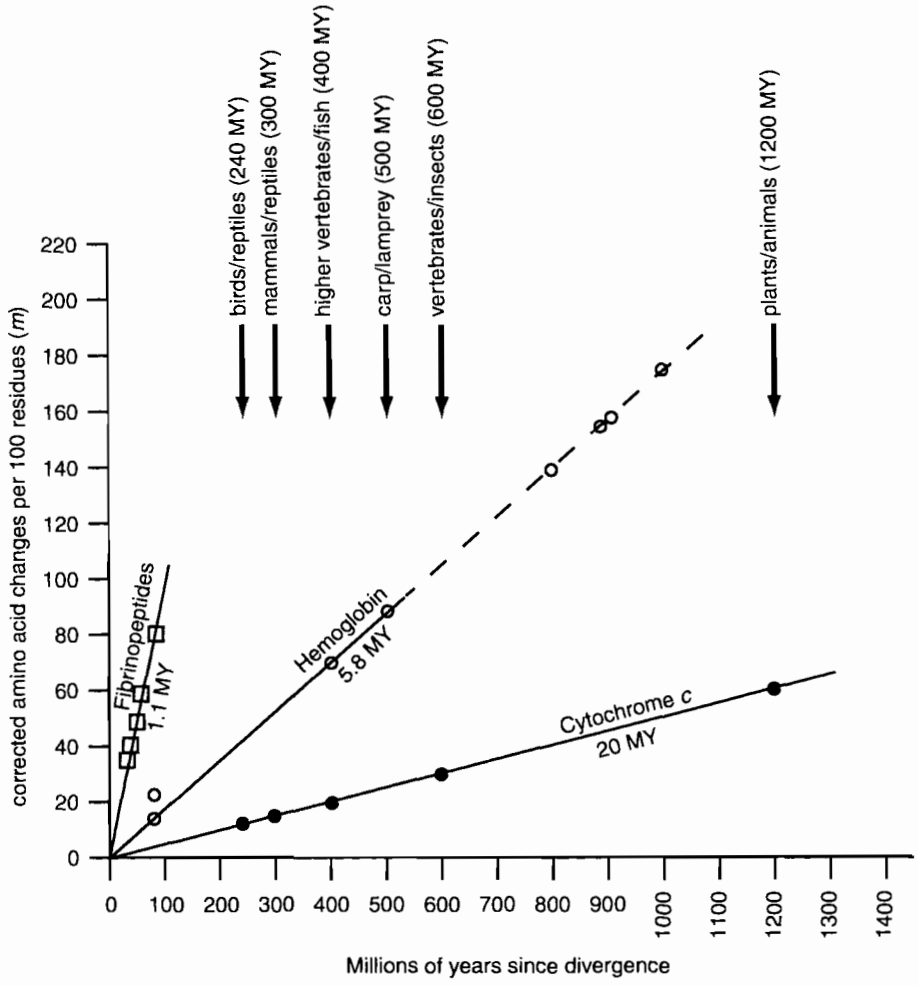


FIGURE 11.3. *A comparison of the number of amino acid changes that occurs between proteins (y axis) versus the time since the species diverged (x axis) reveals that individual proteins evolve at distinct rates. Some proteins, such as cytochrome c from a variety of organisms, evolve very slowly; others such as hemoglobin evolve at an intermediate rate; and proteins such as fibrinopeptides undergo substitutions rapidly. This behavior is described by the molecular clock hypothesis, proposed by Pauling, Margoliash, and others in the 1960s. The time of divergence of various organisms (arrows) is estimated primarily from fossil evidence. Abbreviation: MY, millions of years in the past. Adapted from Dickerson (1971); some data points and the standard deviation measurements are omitted. Used with permission.*

changes per 100 residues. This correction adjusts for amino acid changes that occur but are not directly observed, such as two or more amino acid changes occurring in the same position (see Fig. 11.11 below).

The results of this plot (Fig. 11.3) allow several conclusions (Dickerson, 1971):

- For each protein, the data lie on a straight line. This suggests that the rate of change of amino acid sequence has remained constant for each protein.
- The average rates of change are distinctly different for each protein. For example, fibrinopeptides evolve with a much higher rate of substitution. The time (in millions of years) for a 1% change in amino acid sequence to occur between two divergent lines of evolution is 20.0 MY for cytochrome *c*, 5.8 MY for hemoglobin, and 1.1 MY for fibrinopeptides.
- The observed variations in rate of change between protein families reflect functional constraints imposed by natural selection.

The rate of amino acid substitution is measured by the number of substitutions per amino acid site per year, λ. Some values for λ are given in Table 11.1. Note that some proteins such as histones and ubiquitin undergo substitutions extraordinarily slowly. For reference, Table 11.2 lists the 20 most conserved proteins present in the *Homo sapiens, Caenorhabditis elegans* (nematode), and *Saccharomyces cerevisiae* (yeast) proteomes as determined by reciprocal BLAST searches.

Note that we say that histones undergo substitutions very slowly, but we do not say that they *mutate* very slowly. Mutation is the biochemical process that results in a change in sequence. For example, a polymerase copies DNA with a particular mutation rate. Substitution is the observed change in nucleic acid or protein sequences (e.g. between various histones). The observed substitutions occur at a rate that reflects both mutation and selection, the process by which characters are selected for (or against) in evolution.

TABLE 11-1 Rates of Amino Acid Substitutions per Amino Acid Site per 10^9 Years ($\lambda \times 10^9$) in Various Proteins

Dayhoff (1978) expressed these rates as accepted point mutations (PAMs) per 100 amino acid residues that are estimated to have occurred in 100 million years of evolution (compare Box 3.3, page 50). Thus the rate of mutation acceptance for serum albumin is 19 PAMs per 100 million years.

Protein	Rate	Protein	Rate
Fibrinopeptides	9.0	Thyrotropin beta chain	0.74
Growth hormone	3.7	Parathyrin	0.73
Immunoglobulin (Ig) kappa chain C region	3.7	Parvalbumin	0.70
Kappa casein	3.3	Trypsin	0.59
Ig gamma chain C region	3.1	Melanotropin beta	0.56
Lutropin beta chain	3.0	Alpha crystallin A chain	0.50
Ig lambda chain C region	2.7	Endorphin	0.48
Lactalbumin	2.7	Cytochrome b_5	0.45
Epidermal growth factor	2.6	Insulin (except Guinea pig and coypu)	0.44
Somatrotropin	2.5	Calcitonin	0.43
Pancreatic ribonuclease	2.1	Neurophysin 2	0.36
Serum albumin	1.9	Plastocyanin	0.35
Phospholipase A2	1.9	Lactate dehydrogenase	0.34
Prolactin	1.7	Adenylate kinase	0.32
Carbonic anhydrase C	1.6	Cytochrome *c*	0.22
Hemoglobin alpha chain	1.2	Troponin C, skeletal muscle	0.15
Hemoglobin beta chain	1.2	Alpha crystallin B chain	0.15
Gastrin	0.98	Glucagon	0.12
Lysozyme	0.98	Glutamate dehydrogenase	0.09
Myoglobin	0.89	Histone H2B	0.09
Amyloid AA	0.87	Histone H2A	0.05
Nerve growth factor	0.85	Histone H3	0.014
Acid proteases	0.84	Ubiquitin	0.010
Myelin basic protein	0.74	Histone H4	0.010

Source: Dayhoff (1978, page 3) as adapted by Nei (1987, p. 50). Used with permission.

A significant implication of the molecular clock hypothesis is that if protein sequences evolve at constant rates, then they can be used to estimate the time that the sequences diverged. In this way phylogenetic relationships can be established between organisms. This is analogous to the dating of geological specimens using radioactive decay. An example of how the molecular clock may be used is given in Box 11.1.

The molecular clock hypothesis does not apply to all proteins, and a variety of exceptions and caveats have been noted:

- The rate of molecular evolution varies among different organisms. For example, some viral sequences tend to change extremely rapidly compared to other life forms.
- The clock varies among different genes (see Table 11.1) and across different parts of an individual gene (see, e.g., Fig. 11.1). The main force guiding the molecular clock is selection. Rodents tend to have a faster molecular clock than primates: Their generation times are shorter, and they have high metabolic rates.

TABLE 11-2 **Twenty Most Conserved Proteins Present in *C. Elegans* (worm), *H. Sapiens*, and *S. Cerevisiae* (Yeast)**

	Pairwise Percent Identity		
Protein	Worm/ Human	Worm/ Yeast	Yeast/ Human
1. H4 histone	99	91	92
2. H3.3 histone	99	89	90
3. Actin B	98	88	89
4. Ubiquitin	98	95	96
5. Calmodulin	96	59	58
6. Tubulin 2	94	75	76
7. Ubiquitin-conjugating enzyme UBC4	93	80	80
8. Clathrin coat associated protein	93	48	48
9. Tubulin	93	73	74
10. ADP ribosylation factor 1	93	77	77
11. Dynein light chain 1	92	51	50
12. GTP-binding nuclear protein RAN	89	82	81
13. Ser/Thr protein phosphatase PP1γ	89	84	85
14. Ser/Thr protein phosphatase PP2β	89	74	76
15. Ubiquitin-conjugating enzyme UBE2N	88	67	70
16. Histone H2A.Z	88	69	69
17. Histone H2A.2	87	79	76
18. DIM1P homolog	86	61	65
19. G25K GTP-binding protein	86	76	80
20. 40S ribosomal protein S15A	86	76	77

Note: The data are adapted from Peer Bork and colleagues (Copley et al., 1999), who performed reciprocal BLAST searches against the completed proteomes of these three organisms, as available at the time. As we will see in Chapter 16, genome and proteome annotation change substantially over time. Used with permission.

- The clock is only applicable when a gene in question retains its function over evolutionary time. Genes may become nonfunctional (e.g., pseudogenes) or they may change function after duplication. This may lead to rapid changes in nucleotide sequence.
- The rate of evolution is variable, sometimes accelerating after gene duplication occurs. For example, after gene duplication generated α- and β-hemoglobins, high rates of amino acid substitution occurred that presumably altered the function of the protein.

We will discuss the duplication of an entire genome, followed by subsequent, rapid mutation and gene loss, in Chapter 15 (on the yeast *S. cerevisiae*).

Despite these issues, the molecular clock hypothesis has proven useful and valid in the majority of cases to which it is applied. You may want to estimate the rate of nucleotide substitution for a gene you are studying.

Neutral Theory of Molecular Evolution

There is a tremendous amount of DNA polymorphism in all species that is difficult to account for by conventional natural selection. We will examine this throughout the tree of life in the last third of this book. In Chapter 18, we will examine single nucleotide polymorphisms (SNPs), an extremely common form of polymorphism.

BOX 11-1
Rate of Nucleotide Substitution *r* and Time of Divergence *T*

As an example of how a molecular clock is used, consider the rate of substitution r for any given protein or DNA region for a pair of species in which the time of divergence can be established based upon fossil (paleontological) data. For the α chain of hemoglobin there are 0.56×10^{-9} nonsynonymous nucleotide substitutions per site per year (Gravr and Li, 2000). The α-globins from rat and human differ by 0.093 nonsynonymous substitutions per site. Synonymous substitutions are those that do not result in a codon change that specifies a different amino acid. For example, the codons CCU, CCC, CCA, and CCG each specify the amino acid proline, so any nucleotide change at the third position represents a synonymous substitution. Nonsynonymous substitutions do result in the specification of a different amino acid. For example, any nucleotide change in the codon ATG results in a change from the specified amino acid methionine to something else. For such a codon, it would not make sense to refer to the synonymous substitution rate, since there are no such sites.

The rate of nucleotide substitution r is the number of substitutions K between the two homologous sequences divided by $2T$, where T is the time of divergence between the two sequences. The divergence time for rat and human lineages can then estimated (Graur and Li, 2000, p.140):

$$r = \frac{K}{2T} \tag{11.3}$$

$$T = \frac{0.093}{(2)(0.56 \times 10^{-9})} = 80 \text{ million years ago (MYA)} \tag{11.4}$$

Darwin's theory of evolution suggests that, at the phenotypic level, traits that enhance survival are selected. At the molecular level, a conventional evolutionary point of view is that positive selection also operates on polymorphism in DNA sequences. For example, a gene encoding an enzyme may duplicate (see Chapter 15 and 16), and then subsequent nucleotide changes may allow one of the duplicated genes to encode an enzyme with a novel function that is advantageous and hence selected for.

Motoo Kimura (1968, 1983) proposed a different model to explain evolution at the DNA level. According to his neutral theory, the main cause of evolutionary change (or variability) at the molecular level is random drift of mutant alleles that are selectively neutral or nearly neutral. The fate of these mutant alleles is determined by random genetic drift. Under this model, positive Darwinian selection plays an extremely limited role. As an example, the divergent C peptide of the insulin molecule (Figure 11.1b) changes rapidly according to the neutral mutation rate, while the A and B chains are under selective constraint and evolve more slowly.

While the tree of life provides an appealing metaphor, evolution is not predicated on there necessarily being a single tree. Instead, evolution is based on a process of mutation and selection. We will see in Chapter 14 that genes can be laterally transferred between species, complicating the ways organisms can acquire genes and traits.

Goals of Molecular Phylogeny

All life forms share a common origin and are part of the tree of life. More than 99% of all species that ever lived are extinct (Wilson, 1992). Of the extant species, closely related organisms are descended from more recent common ancestors than distantly related organisms. In principle, there may be one single tree of life that accurately describes the evolution of species. The object of phylogeny is to deduce the correct trees for all species of life. Historically, phylogenetic analyses were based upon easily observable features such as the presence or absence of wings or a spinal cord. More recently, phylogenetic analyses also rely on molecular sequence data that

define families of genes and proteins. Another object of phylogeny is to infer or estimate the time of divergence between organisms since the time they last shared a common ancestor.

The tree of life has three major branches: bacteria, archaea, and eukaryotes. We will explore the global tree in Chapter 12. In this chapter we will address the topic of phylogenetic trees that are used to assess the relationships of homologous proteins (or homologous nucleic acid sequences) in a family. Any group of homologous proteins (or nucleic acid sequences) can be depicted in a phylogenetic tree.

A *true tree* depicts the actual, historical events that occurred in evolution. It is essentially impossible to generate a true tree. Instead, we generate *inferred trees,* which depict a hypothesized version of the historical events. Such trees describe a series of evolutionary events that are inferred from the available data, based on some model.

In Chapter 3, we defined two proteins as homologous if they share a common ancestor. You may perform a BLAST search and observe several proteins with high scores (low expect values) and simply view these database matches as related proteins that possibly have related function. However, it is also useful to view orthologs and paralogs in an evolutionary context. We have applied a variety of approaches to study the relations of proteins: pairwise alignment using Dayhoff's scoring matrices (Chapter 3), BLAST searching (Chapters 4 and 5), the identification of related protein folds (Chapters 8 and 9), and multiple sequence alignment (Chapter 10). All these approaches rely on evolutionary models to account for the observed similarities and differences between molecular sequences:

In 1973 Theodosius Dobzhansky wrote an article entitled "Nothing in biology makes sense except in the light of evolution."

- Dayhoff et al. (1978, p. 345) introduce scoring matrices in explicit evolutionary terms: "An accepted point mutation in a protein is a replacement of one amino acid by another, accepted by natural selection. It is the result of two distinct processes: the first is the occurrence of a mutation in the portion of the gene template producing one amino acid of a protein; the second is the acceptance of the mutation by the species as the new predominant form. To be accepted, the new amino acid usually must function in a way similar to the old one: chemical and physical similarities are found between the amino acids that are observed to interchange frequently."Dayhoff et al. compare observed amino acid sequences from two proteins not with each other but with their inferred ancestor obtained from phylogenetic trees.
- Feng and Doolittle (1987, p. 351) used the Needleman and Wunsch pairwise alignment progressively "to achieve the multiple alignment of a set of protein sequences and to construct an evolutionary tree depicting their relationship. The sequences are assumed a priori to share a common ancestor, and the trees are constructed from different matrices derived directly from the multiple alignment. The thrust of the method involves putting more trust in the comparison of recently diverged sequences than in those evolved in the distant past."
- In our description of protein families, we provided the example of the Pfam JalView tool that allows distance information from the multiple sequence alignment of any Pfam family to be depicted as a tree (Fig. 10.17).

In this chapter, we will use multiple sequence alignments of protein (or DNA or RNA) to generate phylogenetic trees. These trees provide a visualization of the evolutionary history of molecular sequences.

MOLECULAR PHYLOGENY: NOMENCLATURE OF TREES

Molecular phylogeny is the study of the evolutionary relationships among organisms or among molecules using the techniques of molecular biology. Many other

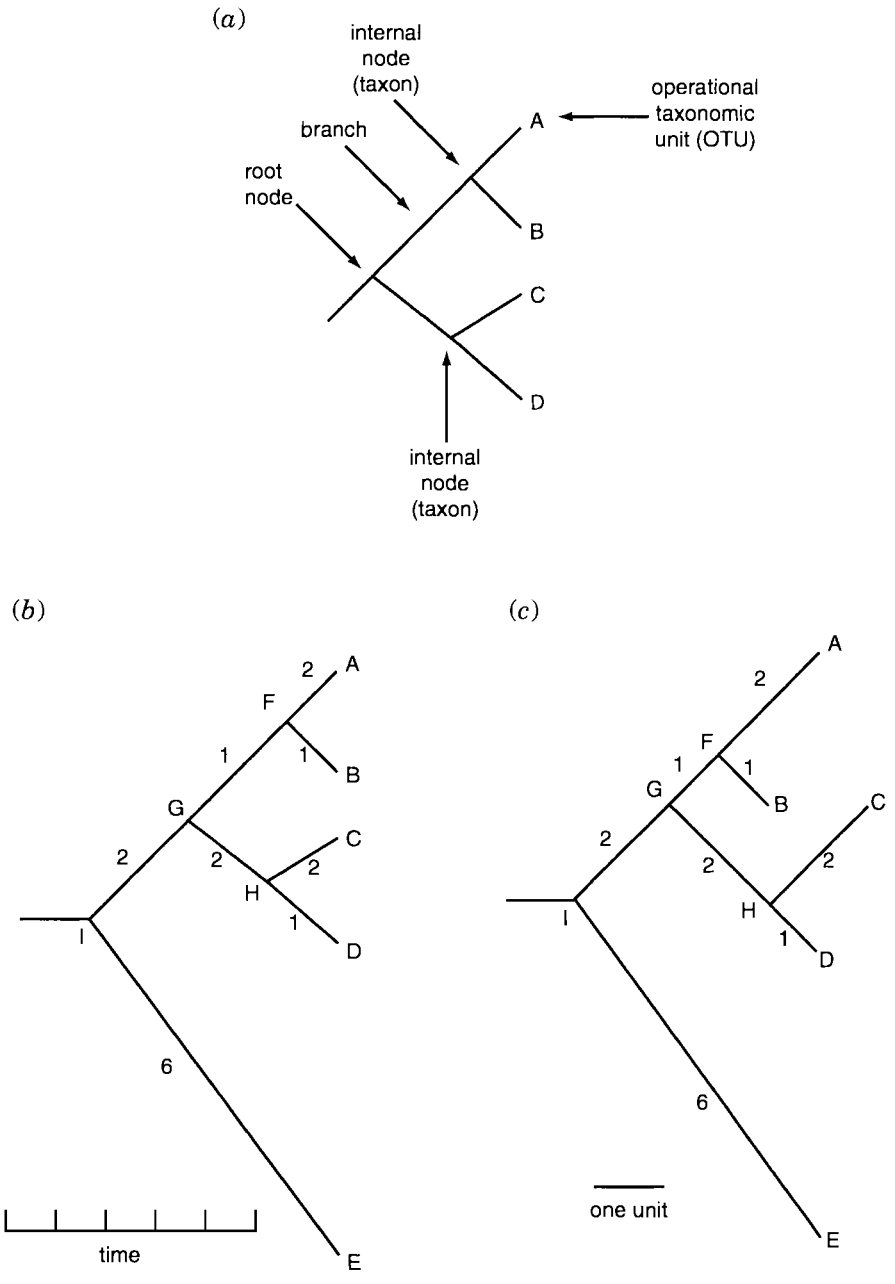


FIGURE 11.4. (a) *Phylogenetic trees contain nodes and branches. A node may be external, internal, or at the root of a tree (the root is defined below). A branch connects two nodes. The nodes represent taxa or taxonomic units; the taxa that provide observable features, such as existing protein sequences or morphological features, are called operational taxonomic units (OTUs). Phylogenetic trees may be* (b) *unscaled or* (c) *scaled. In an unscaled tree, the branch lengths are not proportional to the number of amino acid (or nucleotide) changes. For example, note that branches FA (two units) and FB (one unit) have the same apparent length. Here, operational taxonomic units (ABCDE) are neatly aligned in a column at the tips of the tree. The* x *axis represents time (in units such as millions of years). In the scaled tree in* (b), *the branch lengths are proportional to the number of substitutions. With this topology it is much easier to visualize the relatedness of proteins (or genes) in the tree.*

The topology of a tree defines the relationships of the proteins (or other objects) that are represented in the tree. For example, the topology shows the common ancestor of two homologous protein sequences. The branch lengths sometimes (but not always) reflect the degree of relatedness of the objects in the tree.

techniques are used to study evolution, including morphology, anatomy, paleontology, and physiology. We will focus on phylogenetic trees using molecular sequence data. We begin with an explanation of the nomenclature used to describe trees. There are two main kinds of information inherent in any phylogenetic tree: the topology and the branch lengths. It is necessary to introduce a variety of terms that are used to characterize trees.

Let us first define the main parts of a tree and the main types of trees. A phylogenetic tree is a graph composed of branches and nodes (Fig. 11.4*a*). Only one branch (also called an edge) connects any two nodes. The nodes represent the taxonomic units (taxa or taxons); the node (from the Latin for "knot") is the intersection or terminating point of two or more branches. For us, taxa will typically be protein

sequences. An operational taxonomic unit (OTU) is an extant taxon present at an external node, or leaf; the OTUs are the available nucleic acid or protein sequences that we are analyzing in a tree.

In our discussion of trees, it is assumed that the raw data may consist of DNA, RNA, or protein sequence data. These data are presented as a multiple sequence alignment.

Consider the two trees in Figs. 11.4*b,c*. Each tree consists of five OTUs (labeled A, B, C, D, and E). These five OTUs define five external nodes. In addition, there are internal nodes at positions F, G, H, and I. Each internal node represents an inferred ancestor of the OTUs. Imagine that the tree is of five lipocalins, and A and B correspond to human and rat RBP. The internal node that connects to A and B represents an ancestral sequence that existed in an organism that predated the divergence of primates and rodents some 80 MYA.

Branches define the topology of the tree, that is, the relationships among the taxa in terms of ancestry. In some trees, the branch length represents the number of amino acid changes that have occurred in that branch. Branches of a tree are also called edges. In the trees of Figure 11.4, the branches leading to each of the OTUs are called external branches (or peripheral branches). The branches leading to F, G, H, and I are called internal branches.

In the example of Figure 11.4*b*, the branches are unscaled. This implies that they are not proportional to the number of changes. This form of presenting a tree (called a cladogram) has the advantage of aligning the OTUs neatly in a vertical column. This may be especially useful if the tree has many dozens of OTUs. Also, this format allows the nodes to be placed along a time scale describing when the divergence event is estimated to have occurred.

In the tree of Figure 11.4*c*, the same raw data are used to generate the tree, but the branch lengths are now scaled. Thus, the branch lengths are proportional to the number of amino acid (or nucleotide) changes that occurred between the sequences. This format (called a phylogram) has the helpful feature of conveying a clear visual idea of the relatedness of different proteins within the tree.

An internal node is bifurcating if it has only two immediate descendant lineages (branches). Bifurcating trees are also called binary or dichotomous; any branch that divides splits into two daughter branches. A tree is multifurcating if it has a node with more than two immediate descendants (Fig. 11.5, ABC).

A multifurcation is also called a polytomy. Multifurcating trees are by definition nonbinary.

A clade is a group of all the taxa that have been derived from a common ancestor plus the common ancestor itself. A clade is also called a monophyletic group. In our context, a clade is a set of proteins that form a group within a tree. In the example of either tree in Figs. 11.4*b,c*, C, D, and H form a clade, but B is not a member of this clade. A larger clade is defined by C, D, H, A, B, F, and G. The OTU labeled E is not a member of this larger clade. The taxonomic group ABF that shares a common ancestor (G) with another taxonomic group (CDH) is paraphyletic.

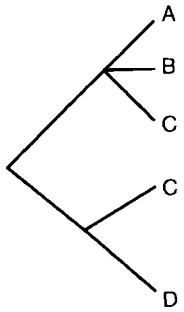


FIGURE 11.5. *A phylogenetic tree is said to be multifurcating or polytomous if it has a node with three or more branches (see the node leading to taxa ABC). It is common to make a multifurcating tree when the available data do not provide enough information to define a tree with only bifurcating nodes.*

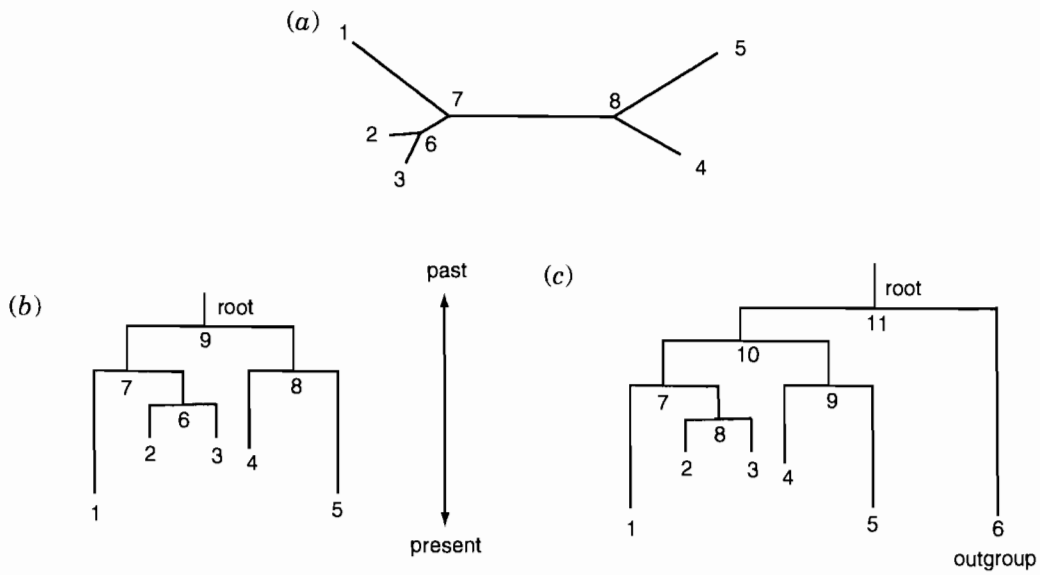


FIGURE 11.6. *A phylogenetic tree may be* (a) *unrooted or* (b) *rooted. The same raw data are used to generate each type of tree in* (a) *and* (b). *The placement of a root implies a hypothesis about the common ancestor of all members of the tree. When this information is not known, an unrooted tree may be more appropriate. Rooting of a tree may be accomplished in two ways. In midpoint rooting the longest branch [here, the branch connecting nodes 7 and 8 in* (b)*] may be used to define the most likely place to add a root.* (c) *A single taxon from a phylogenetically distant organism is added to the data set (taxon 6) and used to define an outgroup in a new tree. For example, an invertebrate lipocalin may form an outgroup relative to a series of mammalian lipocalins.*

Tree Roots

A phylogenetic tree has a root representing a most recent common ancestor of all the sequences. Often this root is not known today, and some tree-making algorithms do not provide conjectures about placement of a root. The alternative to a rooted tree is an unrooted tree. An unrooted tree specifies the relationships among the OTUs. However, it does not define the evolutionary path completely or make assumptions about common ancestors. Figure 11.6 shows a binary tree with five OTUs that is either unrooted (Fig. 11.6*a*) or rooted (Fig. 11.6*b*). The OTUs (extant taxons, leaves) are numbered 1–5. Some OTUs can be swapped (exchanged) without altering the topology of the tree, such as 4 and 5 in either tree. But others cannot be swapped, such as 1 and 2. Note that in the unrooted tree the direction of time is undetermined.

Keep in mind that a bacterial lipocalin may have emerged in the immediate past or as long as several billion years ago. Regardless of its age on the planet, its rate of evolution is not necessarily faster or slower than that of any other lipocalin.

The principal way to root a tree is to specify an outgroup. In Figure 11.6*c*, imagine that sequences 1–5 are mammalian RBP orthologs and that the sequence of a homologous bacterial or invertebrate protein (OTU 6) is obtained. This invertebrate sequence is clearly derived from a common ancestor that predates the appearance of all the other OTUs. Thus it can be used to define the location of the root.

You can select an OTU in order to place a root by identifying the most closely related outgroup. A second way to place a root is through midpoint rooting. Here, the longest branch is determined (such as the branch between internal nodes 7 and 8 in Fig. 11.6*a*). This longest branch is presumed to be the most reasonable site for a root.

Enumerating Trees

The number of possible trees to describe the relationships of a dozen protein sequences is staggeringly large. It is important to know the number of possible trees

BOX 11-2
Number of Rooted and Unrooted Trees

The number of bifurcating unrooted trees (N_U) for n OTUs ($n \geq 3$) is given by Cavalli-Storza and Edwards (1967):

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

The number of bifurcating rooted trees (N_R) for n OTUs ($n \geq 2$) is

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

For example, for four OTUs, N_R equals $(8-3)!/(2^2)(2)! = 5!/8 = 15$. The number of possible rooted and unrooted trees (up to 50 OTUs) is as follows. The values were calculated using MatLab software (MathWorks).

Number of OTUs	Number of Rooted Trees	Number of Unrooted Trees
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10,395	945
8	135,135	10,395
9	2,027,025	135,135
10	34,489,707	2,027,025
15	213,458,046,676,875	8×10^{12}
20	8×10^{21}	2×10^{20}
50	2.8×10^{76}	3×10^{74}

To give a sense of the immense number of possible trees corresponding to just a few dozen taxa, there are on the order of 10^{79} protons in the universe.

for any tree you are making. There is only one "true" tree representing the evolutionary path by which molecular sequences (or even species) evolved. The number of potential trees is useful in deciding which tree-making algorithms to apply.

The number of possible rooted and unrooted trees is described in Box 11.2. For two OTUs, there is only 1 tree possible. For three taxa, it is possible to construct either 1 unrooted tree or 3 different rooted trees (Fig. 11.7). For four taxa, the number of possible trees rises to 3 unrooted trees or 15 rooted trees (Fig. 11.8).

An important practical limit is reached at 10 sequences, for which there are over 2 million possible unrooted trees and 34 million rooted trees. It is generally necessary to use a heuristic algorithm to evaluate the robustness of more than this many trees. In contrast, for about ten texa (or fewer) it is possible for a standard desktop computer to perform exhaustive searches for which all possible trees are evaluated.

Some phylogeny projects involve the generation of trees for thousands of taxa. See the Deep Green plant project (▸ http://ucjeps.berkeley.edu/bryolab/greenplantpage.html). The Ribosomal Database (▸ http://rdp.cme.msu.edu/html/) includes an analysis of over 50,000 aligned sequences. For typical analyses, you may analyze several dozen taxa. If you want to make a phylogenetic tree with the lipocalins that are currently in Pfam (version 9.0), you could use the 58 proteins available in the seed alignment or all 396 proteins available in the full alignment.

Species Trees Versus Gene/Protein Trees

Species evolve and proteins (and genes) evolve. The analysis of protein evolution can be complicated by the time that two species diverged. Speciation, the process

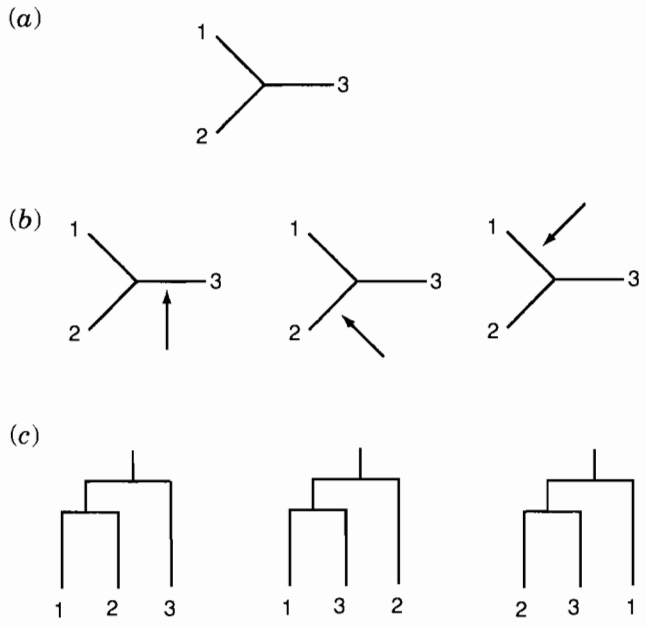


FIGURE 11.7. (a) *For three operational taxonomic units (such as three aligned protein sequences 1–3), there is one possible unrooted tree.* (b) *Any of these edges may be used to select a root, from which* (c) *three corresponding rooted trees are possible.*

A heuristic algorithm will explore a subset of all possible trees, discarding vast numbers of trees that have a topology that is unlikely to be useful. Using heuristic algorithms (Box 2.2), it is possible to create phylogenetic trees having hundreds of protein (or DNA) sequences. As an example of how a heuristic algorithm works, consider a data set in which the algorithm seeks a tree with the shortest total branch lengths (i.e., the most parsimonious tree). This search occurs without evaluating all possible trees, but instead by performing a series of rearrangements of the topology. Once a tree with a particular score is obtained, the algorithm can establish that score as an upper limit and discard all trees for which rearrangements are unlikely to yield a shorter tree.

by which two new species are created from a single ancestral species, occurs when the species become reproductively isolated (Fig. 11.9). In a species tree, an internal node represents a speciation event. In a gene tree (or protein tree), an internal node represents the divergence of an ancestral gene into two new genes with distinct sequences.

In a genetically polymorphic population, gene duplication events may occur before or after speciation. A protein (or gene) tree differs from a species tree in two ways (Graur and Li, 2000): (1) The divergence of two genes from two species may have predated the speciation event. This may cause overestimation of branch lengths in a phylogenetic analysis. (2) The topology of the gene tree may differ from that of the species tree. In particular, it may be difficult to reconstruct a species tree from a gene tree. A molecular clock may be applied to a gene tree in order to date the time of gene divergence, but it cannot be assumed that this is also the time that speciation occurred.

Reconstructing a phylogenetic tree based upon a single protein (or gene) can thus give complicated results. For this reason, many researchers construct trees from a variety of distinct protein (or gene) families in order to assess the relationships of different species. Another strategy that has been adopted is to generate concatenated protein (or DNA) sequences. For example, Baldauf et al. (2000) used four concatenated protein sequences to create a comprehensive phylogenetic tree of eukaryotes (Fig. 16.1).

Four Stages of Phylogenetic Analysis

Molecular phylogenetic analyses can be divided into four stages: (1) selection of sequences for analysis, (2) multiple sequence alignment of homologous protein or nucleic acid sequences, (3) tree building, and (4) tree evaluation.

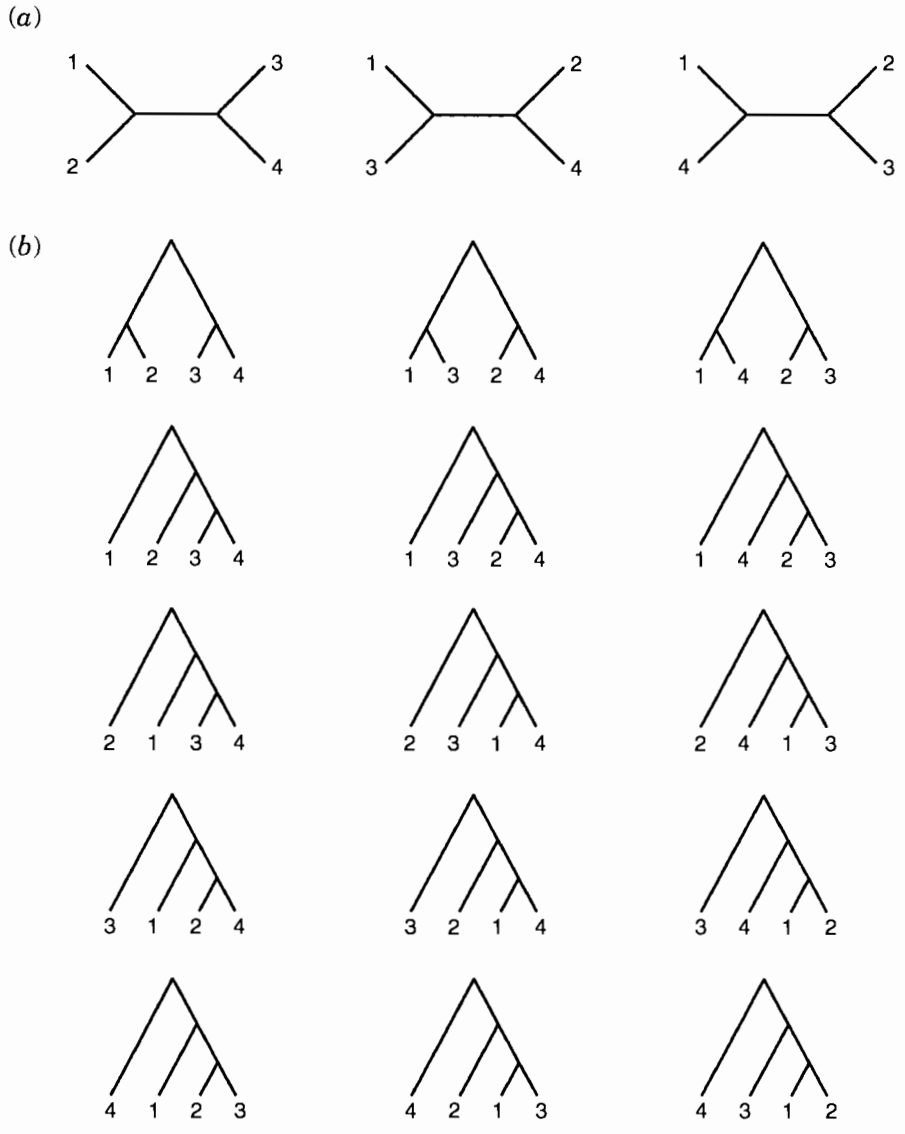


FIGURE 11.8. *For four operational taxonomic units (such as four aligned protein sequences 1–4), there are* (a) *3 possible unrooted trees and* (b) *15 possible rooted trees. Only one of these is a true tree in which the topology accurately describes the evolutionary process by which these sequences evolved.*

Stage 1: Molecular Phylogeny Can Be Performed with DNA, RNA, or Protein

When you generate a phylogenetic tree using molecular sequence data, you can use DNA, RNA, or protein sequences. In one common scenario, you may want to evaluate the relationship of a group of molecules such as lipocalins. The choice of whether to study protein or DNA depends in part on the question you are asking. In some cases, protein studies are preferable; you may prefer to study a multiple sequence alignment of proteins, or the lower rate of substitutions in protein relative to DNA may make protein studies more appropriate for comparisons across widely divergent species. In many other cases, studying DNA is more informative than protein. There are several reasons for this.

- The portion of DNA that codes for a protein can have both synonymous and nonsynonymous substitutions. This is depicted for the alignment of three RBPs at their 5′ ends (amino termini of the proteins) and 3′ ends (carboxy termini of the proteins) (Fig. 11.10; see arrows 11–14 and 21–34).

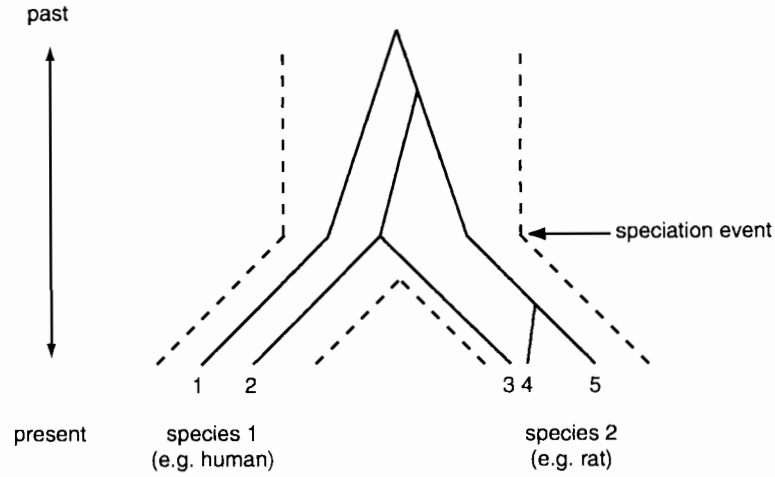


FIGURE 11.9. *A species tree and a protein (or gene) tree can have a complex relationship. A speciation event, such as the divergence of the lineage that generated modern humans and rodents, can be dated to a specific time (e.g., 80 MYA). When speciation occurs, the species become reproductively isolated from one another. This event is represented by dotted lines (see horizontal arrow). Phylogenetic analysis of a specific group of homologous proteins is complicated by the fact that a gene duplication could have preceded or followed the speciation event. In essentially all phylogenetic analyses, the extant proteins (OTUs) are sequences from organisms that are alive today. It is necessary to reconstruct the history of the protein family as well as the history of each species. In the above example, there are two human paralogs and three rat paralogs. Proteins 1 and 5 diverged at a time that greatly predates the divergence of the two species. Proteins 2 and 3 diverged at a time that matches the date of species divergence. Proteins 4 and 5 diverged recently, after the time of species divergence. It is possible to reconstruct both species trees and protein (or gene) trees. Modified from Nei (1987) and Graur and Li (2000).*

- Comparison of the rates of nonsynonymous substitution ($\hat{d}_N$) versus synonymous substitution ($\hat{d}_S$) may reveal evidence of positive or negative selection. If $\hat{d}_S$ is greater than $\hat{d}_N$, this suggests that the DNA sequence is under negative or purifying selection. Negative selection is selection that limits change in a corresponding amino acid sequence; this occurs when some aspect of the structure and/or function of a protein is critical and cannot tolerate substitutions. When $\hat{d}_N$ is greater than $\hat{d}_S$, this suggests that positive selection occurs. An example of positive selection is with a duplicated gene that is under pressure to evolve new functions.
- Substitutions in DNA include those that are directly observed in an alignment, such as single-nucleotide substitutions, sequential substitutions, and coincidental substitutions (depicted in Figs. 11.11*a,c*). By analyzing two sequences with reference to an ancestral sequence (Fig. 11.11*b*), it is possible to infer a great deal of information about mutations that do not appear in a direct comparison of two (or more) sequences. These mutational processes include parallel substitutions, convergent substitutions, and back substitutions (Fig. 11.11*c*).
- Noncoding regions (such as the 5′ and 3′ untranslated regions of genes, or introns) may be analyzed using molecular phylogeny (Fig. 11.10). In many cases, there is little evolutionary pressure to conserve the nucleotide sequence, and these regions may vary greatly. That is, the nucleotide substitution rate equals the neutral mutation rate. In other cases there is tremendous nucleotide conservation, perhaps because of the presence of a regulatory element such as a transcription factor binding motif.

We encountered the idea that more mutational events occur than can be directly observed when we examined PAM matrices (Chapter 3). There, we saw that two proteins of length 100 that share 50% amino acid identity have sustained an average of 80 changes (Fig. 3.19).

Synapomorphy is defined as a character state that is shared by several taxa. Homoplasy is defined as a character state that arises independently (e.g., through convergent substitutions or back substitutions) but is not derived from a common ancestor (i.e. is not homologous). See Graur and Li (2000).

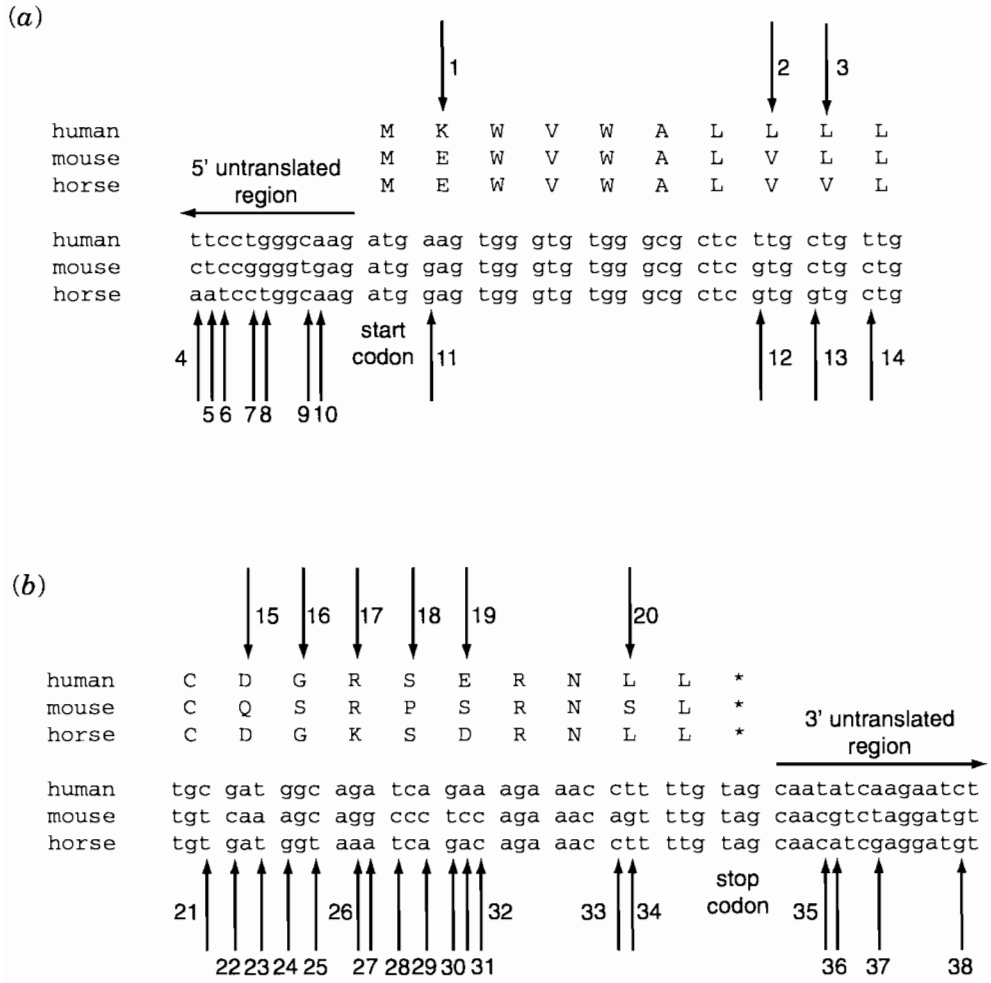


FIGURE 11.10. *Phylogenetic trees can be constructed using DNA, RNA, or protein sequence data. Often, the DNA sequence is more informative than protein in phylogenetic analysis. As an example, the sequences of retinol-binding protein from three species are aligned at the 5′ end of the DNA [with the corresponding amino termini of the proteins in* (a)*] and the 3′ end of the DNA [carboxy termini of the proteins in* (b)*]. In the 5′ and 3′ untranslated regions, where no protein is encoded, there is typically less selective pressure to maintain particular nucleotide residues. (Some regulatory elements may be highly conserved.) Here, 7 of 12 nucleotide positions vary at the 5′ end (arrows 4–10) and at the 3′ end several nucleotides vary (arrows 35–38). Within the protein-coding region, there are variant amino acid residues at 3 of the first 10 positions (arrows 1–3) and 6 of the last 10 positions (arrows 15–20). These variants may be informative in performing phylogeny. However, there is an even greater number of informative nucleotide changes, restricting our attention to the coding region: In addition to the nonsynonymous substitutions (causing amino acid changes; arrows 11–13, 22–24, 26, 28, 30–34), there are a few synonymous changes (not causing a different amino acid to be specified; arrows 14, 21, 25, 27, 29). The RBP sequences are from human (GenBank accession NM_006734), mouse (XM_129259), and horse (U21208).*

- Pseudogenes have been studied using molecular phylogeny. By definition, these do not encode functional proteins (see Chapter 12).
- The rate of transitions and transversions can be evaluated. In some cases, transitions occur 10 times more frequently (Moritz and Hillis, 1996).

While the analysis of DNA offers many advantages, it is often preferable to study proteins for phylogenetic analysis. The evolutionary distance between two organisms may be so great that any DNA sequences are saturated. That is, at many

A transition is a nucleotide substitution between two purines (A to G or G to A) or between two pyrimidines (C to T or T to C). A transversion is the substitution between a purine and a pyrimidine (e.g., A to C, C to A, G to T; there are eight possible transversions). The International Union of Pure and Applied Chemistry (IUPAC; ▸http://www.iupac.org) defines many symbols commonly used in science. The abbreviations of the four nucleotides are adenine (A), cytosine (C), guanine (G), and thymine (T). Additional abbreviations are for an unspecified or unknown nucleotide (N), an unspecified purine nucleotide (R), and an unspecified pyrimidine nucleotide (Y).

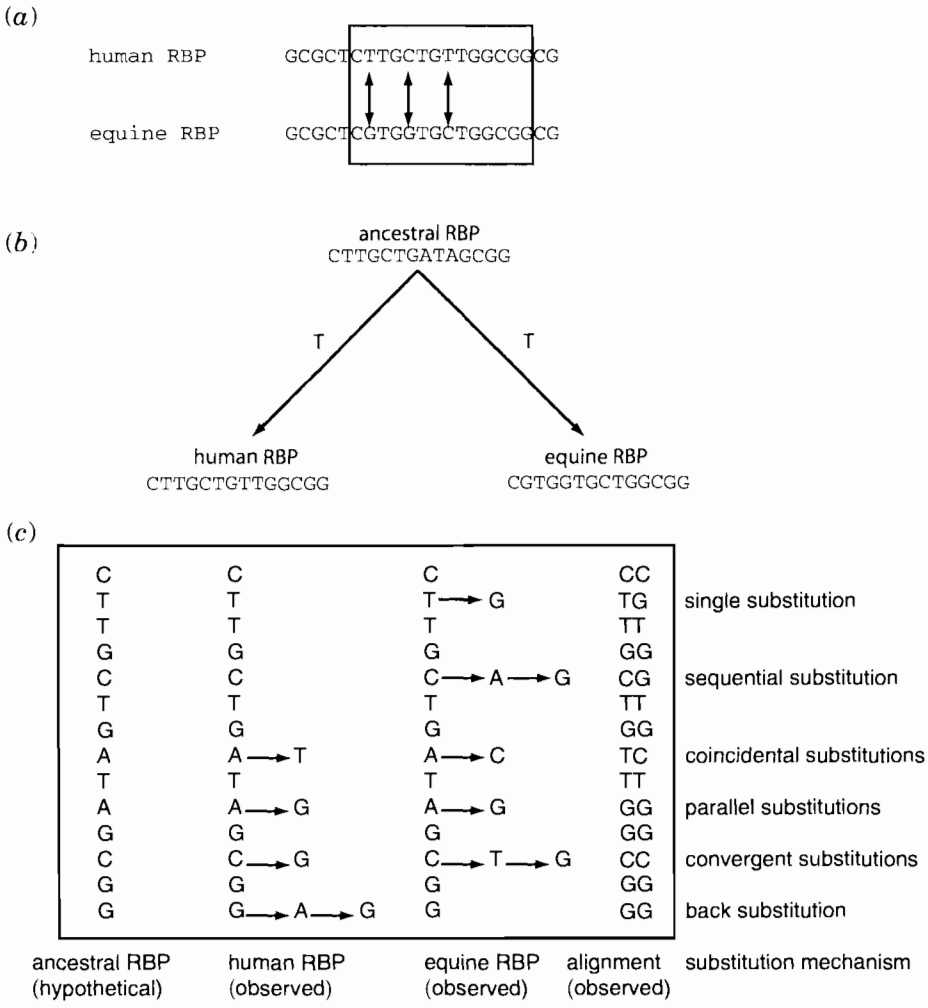


FIGURE 11.11. *Multiple types of mutations occur in sequences.* (a) *A portion of the coding sequence of human and equine RBP, including three observed mismatches. Many more than three mutations may have occurred in this region.* (b) *There is a hypothetical, ancestral RBP sequence from which human and equine RBP diverged in the past at time* T. (c) *Comparison of the hypothetical ancestral sequence with the observed human and equine sequences illustrates several mutational mechanisms. Single-nucleotide substitution, sequential substitution, and coincidental substitution all could account for observed mutations (red-colored nucleotides). Parallel, convergent, and back substitutions all could occur without producing an observed mismatch. In this example, 3 mutations are observed while 12 mutations actually occurred.* [(a, c) *Adapted from Graur and Li (2000).*] *Used with permission.*

We will show how the entire genome of a fungus duplicated (Chapter 15). The evidence for this consisted of Blastp searches of all *Saccharomyces cerevisiae* proteins against each other, resulting in the detection of conserved blocks of sequence from various chromosomes (see Figure 15.9). Here, Blastn searches would not have been sensitive enough to reveal the homology between different chromosomes.

sites all the possible nucleotide changes may occur (even multiple times), so that phylogenetic signal is lost. Proteins have 20 states (amino acids) instead of only four states for DNA, so there is a stronger phylogenetic signal. For closely related sequences, such as mouse versus rat RBP, DNA-based phylogeny might be more appropriate than protein studies, because the phylogenetic signal is relatively stronger for DNA (see Figures 11.10 and 11.11).

Whether nucleotides or amino acids are selected for phylogenetic analysis, the effects of character changes can be defined. An unordered character is a nucleotide or amino acid that changes to another character in one step. An ordered character is one that must pass through one or more intermediate states before it changes to a different character. Partially ordered characters have a variable or indeterminate number of states between the starting value and the ending value. Nucleotides are unordered characters: any one nucleotide can change to any other in one step (Fig. 11.12*a*). Amino acids are partially ordered. If you inspect the genetic code, you will see that some amino acids can change to a different amino acid in a single step of one nucleotide substitution, while other amino acid changes require two or even three nucleotide mutations (Fig. 11.12*b*).

(*a*)

	A	C	T	G
A	0	1	1	1
C	1	0	1	1
T	1	1	0	1
G	1	1	1	0

(*b*)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	2	1	1	2	1	2	2	2	2	2	2	1	2	2	1	1	1	2	2
C		0	2	3	1	1	2	2	3	2	3	2	2	3	1	1	2	2	1	1
D			0	1	2	1	1	2	2	2	3	1	2	2	2	2	2	1	3	1
E				0	3	1	2	2	1	2	2	2	2	1	2	2	2	1	2	2
F					0	2	2	1	3	1	2	2	2	3	2	1	2	1	2	1
G						0	2	2	2	2	2	2	2	2	1	1	2	1	1	2
H							0	2	2	1	3	1	1	1	1	2	2	2	3	1
I								0	1	1	1	1	2	2	1	1	1	1	3	2
K									0	2	1	1	2	1	1	2	1	2	2	2
L										0	1	2	1	1	1	1	2	1	1	2
M											0	2	2	2	1	2	1	1	2	3
N												0	2	2	2	1	1	2	3	1
P													0	1	1	1	1	2	2	2
Q														0	1	2	2	2	2	2
R															0	1	1	2	1	2
S																0	1	2	1	1
T																	0	2	2	2
V																		0	2	2
W																			0	2
Y																				0

FIGURE 11.12. *Step matrices for* (a) *nucleotides or* (b) *amino acids describe the number of steps required to change from one character to another. For the amino acids, between one and three nucleotide mutations are required to change any one residue to another. Adapted from Graur and Li (2000). Used with permission.*

Stage 2: Multiple Sequence Alignment

Multiple sequence alignment (Chapter 10) is a critical step of phylogenetic analysis. In many cases, the alignment of nucleotide or amino acid residues in a column implies that they share a common ancestor (see character-based methods, below). If you misalign a group of sequences, you will still be able to produce a tree. However, it is not likely that the tree will be biologically meaningful. And if you create a multiple alignment of sequences and include a nonhomologous sequence, it will still be incorporated into the phylogenetic tree.

In preparing a multiple sequence alignment for phylogenetic analysis, there are several important considerations in creating and editing the alignment. Let us introduce these ideas by referring to a specific example of 13 orthologous RBPs. We presented a phylogenetic tree of these proteins in Figure 3.2. The multiple sequence alignment from which this tree was generated is partially shown in Figure 11.13. There are several notable features:

1. Carefully inspect the alignment to be sure that all sequences are homologous. It is sometimes possible to identify a sequence that is so distantly related that it is not homologous. You can further test this possibility by performing pairwise alignments (is the expect value significant?) and/or BLAST searches. If the sequence is not apparently homologous, it should be removed from the multiple sequence alignment.

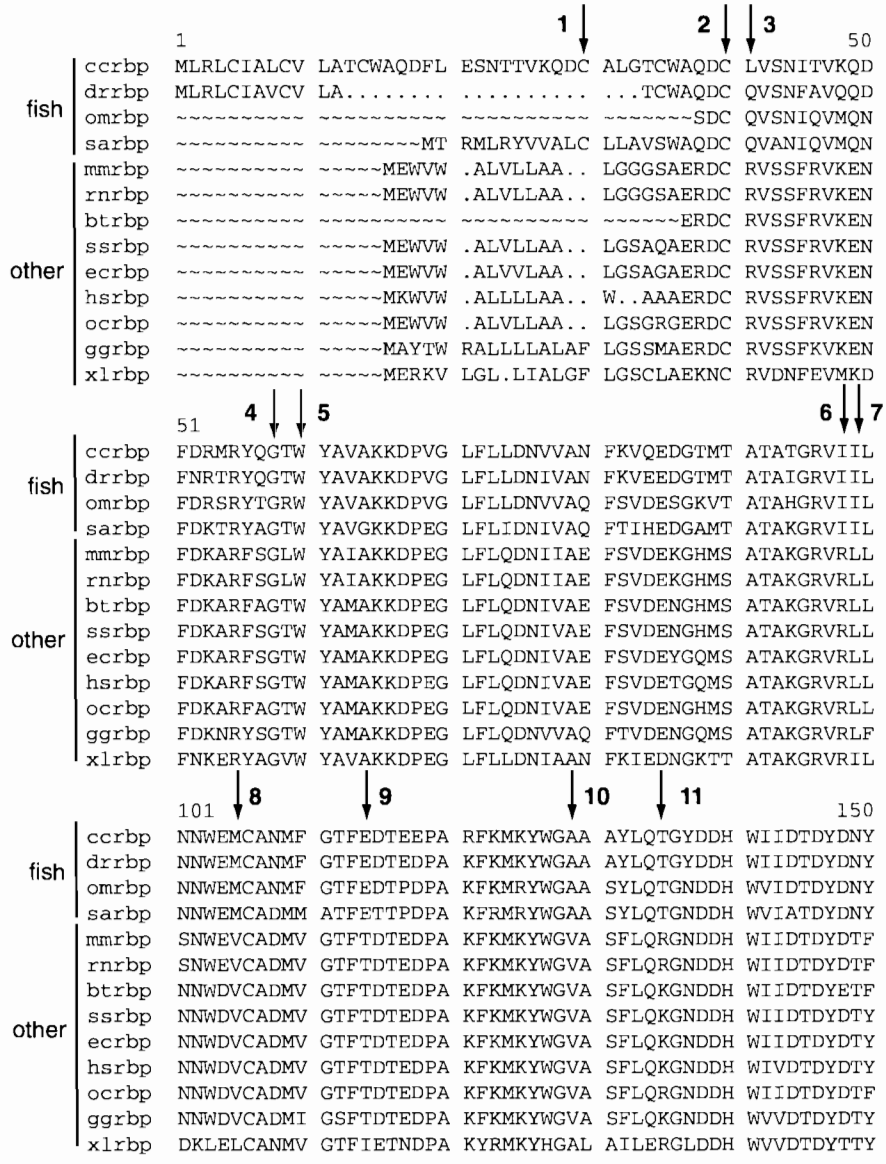


FIGURE 11.13. *We will introduce tree-making approaches with a multiple sequence alignment of 13 orthologous RBPs, made in PileUp (GCG). Only the amino-terminal part of the alignment is shown. Similar data were used in Figure 3.8 to generate the tree shown in Figure 3.2 (see this figure for accession numbers). Four of the sequences are from fish (vertical side bar;* ccrbp *is RBP from* Cyprinus carpio *(carp);* dr, Danio rerio *(zebrafish);* om, Oncorhynchus mykiss *(rainbow trout);* sa, Sparus aurata *(a teleost). One RBP is from an amphibian,* Xenopus laevis (xl). *The other nine are from mouse* (mm), *rat* (rn), *cow* (bt), *pig* (ss), *horse* (ec), *human* (hs), *rabbit* (oc), *and chicken* (gg). *Internal gaps in the alignment (arrow 1) are not easily interpretable by phylogenetic algorithms and could represent either insertions or deletions. Many positions are 100% conserved (e.g., cysteine at arrow 2 and GXW motif at arrows 4 and 5). Amino acids in many other positions are different in fish relative to all other sequences (arrows 3, 6, 8, 11) or occur as a group in fish and the amphibian (arrows 7, 10). A phylogenetic tree provides a visualization of these relationships (Fig. 3.2 and this chapter).*

2. Some multiple sequence alignment programs, such as PileUp and ClustalW, may treat distantly related sequences by aligning them outside the block of other sequences. If necessary, lower the gap creation and/or gap extension penalties to accommodate the distantly related homolog(s) into the multiple sequence alignment.

3. The complete sequence is not known for several of the sequences: there is no apparent start methionine for the *Oncorhynchus mykiss* (rainbow trout) or *Bos taurus* (bovine) RBP. In general, the multiple sequence alignment data used for phylogenetic analyses should be restricted to portions of the proteins (or nucleic acids) that are available for all the taxa being studied.

4. There are occasional internal gaps in this alignment (Fig. 11.13, arrow 1). A gap could represent an insertion in some of the sequences or a deletion in the others. Most phylogeny algorithms are not equipped to evaluate insertions or

deletions (also called indels). Many experts recommend that any column of a multiple sequence alignment that includes a gap in any position should be deleted.

5. In this example, note that there are RBP sequences from four fish (carp, zebrafish, rainbow trout, and a hermaphrodite teleost). Intuitively, we expect these four sequences to be quite different than the other nine sequences, which include mammals (e.g., human), an amphibian (*Xenopus laevis*), and chicken (an Archosauria, the group that includes crocodiles and birds). Thus, although all 13 species having RBP are vertebrates, we expect the fish to be distinctly different than the rest, presumably having shared a common ancestor hundreds of millions of years ago. Indeed, we can see such differences by inspecting the multiple sequence alignment. There are positions in which the amino acid in a particular position differs between fish and the other nine species. Examples are shown in Figure 11.13 (arrows 3, 6, 8, and 11). Other positions are shared in common between fish and the amphibian (arrows 7, 10). Dozens of other positions are entirely conserved among all these proteins (examples are shown at arrows 2, 4, and 5), as expected for a family of closely related proteins. The phylogenetic tree (Fig. 3.2) visualizes these relationships. Any time you inspect a multiple sequence alignment and a tree, you are looking at similar information from different perspectives.

6. As pointed out in Chapter 10, the most important factor in generating an accurate phylogenetic tree is to make a good multiple sequence alignment. This is especially difficult to achieve for poorly conserved proteins, when ambiguities in the positioning of gaps make many significantly different alignments possible. Many authorities recommend removal of segments of the alignment in which gap placement is ambiguous. Arbitrary alignment decisions in such regions may dominate the phylogenetic results.

Stage 3: Tree-Building Methods

There are many ways to build a phylogenetic tree, reviewed in books (Nei, 1987; Graur and Li, 2000; Li, 1997; Maddison and Maddison, 2000; Durbin et al., 2000; Baxevanis and Ouellette, 2001; Clote and Backofen, 2000; Hall, 2001) and articles (Felsenstein, 1996, 1988; Hein, 1990; Nei, 1996; Thornton and DeSalle, 2000). We will consider two principal methods of making trees: distance-based methods and character-based methods.

Distance-based methods begin the construction of a tree by calculating the distances between molecular sequences. There is some distance metric, such as the number of amino acid changes between the sequences, or a distance score (see Box 10.1). Distance-based methods are computationally fast. They involve information about the distance score between the OTUs in a multiple sequence alignment. Thus, a matrix of pairwise scores for all the aligned proteins (or nucleic acid sequences) is used to generate a tree. The main distance-based methods are the unweighted pair group method with arithmetic mean (UPGMA) and neighbor joining (NJ). The Clusters of Orthologous Groups (COG) database (Table 8.9 and Chapter 14) relies on distance-based assignments of gene relatedness.

There are two main character-based methods: maximum parsimony (MP) and maximum likelihood (ML). Parsimony analysis involves a search for the tree with the fewest number of amino acid (or nucleotide) character changes that accounts for the observed differences between protein (or gene) sequences. Parsimony is a more powerful approach than distance to describe the hierarchical relationship of genes

Joe Felsenstein (of the Department of Genome Sciences at the University of Washington) offers a web page with about 200 phylogeny software links (▸http://evolution.genetics.washington.edu/phylip/software.html). His site includes information on his PHYLIP software (the PHYLogeny Inference Package) (see ▸http://evolution.genetics.washington.edu/phylip/general.html). This is one of the most popular and useful programs for phylogenetic analysis, and is freely available for all major computer operating systems.

The HIV Sequence Database at the Los Alamos National Laboratory (discussed in Chapter 13, on viruses) offers a brief online guide to making and interpreting phylogenetic trees (▸http://hiv-web.lanl.gov/content/hiv-db/TREE_TUTORIAL/Tree-tutorial.html). This site includes links to PAUP (discussed below), PHYLIP, and other tree-making programs.

and proteins. The assumption of phylogenetic systematics is that genes exist in a nested hierarchy of relatedness, and this is reflected in a hierarchical distribution of shared characters in the sequences. The most parsimonious tree is supposed to best describe the relationships of proteins (or genes) that are derived from common ancestors.

The word *parsimony* (from the Latin *parcere*, "to spare") refers to simplicity of assumptions in a logical formulation.

We can inspect the multiple sequence alignment in Figure 11.13 to think about the essence of distance-based versus character-based molecular phylogeny. In distance-based approaches, one can calculate the percent amino acid similarity between each pair of proteins in the multiple sequence alignment. Some pairs, such as rat and mouse RBP, are very closely related and will be placed close together in the tree. Others, such as the group of four fish RBPs, are more distant than the other sequences and will be placed farther away on the tree. In a sense, we can look at the sequences in Figure 11.13 horizontally, calculating distance measurements between the entire sequences.

In contrast, character-based approaches regard the multiple sequence alignment from a vertical perspective. In each column of amino acids, what is the simplest (most parsimonious) explanation for how the characters (e.g., amino acids) evolved? The evidence that fish should be placed apart from the other proteins in a tree comes from examining positions in Figure 11.13, as shown in arrows 3, 6, 8, and 11.

Regardless of whether distance-based or character-based approaches are employed, a variety of tree-building programs accept a multiple sequence alignment as input. ReadSeq is a convenient web-based program that translates multiple sequence alignments into formats compatible with most commonly used phylogeny packages. Several ReadSeq servers are listed in Table 11.3.

First released in 1993, ReadSeq was written by Don Gilbert and is in the public domain.

Making Trees Using Distance-Based Methods

There are several ways to measure distances between sequences. The simplest approach is to align pairs of sequences and count the number of differences. The degree of divergence is also called the Hamming distance. For an alignment of length N with n sites at which there are differences, the degree of divergence D is defined as

$$D = \frac{n}{N} \times 100 \tag{11.5}$$

The Hamming distance is simple to calculate, but it ignores a large amount of information about the evolutionary relationships among the sequences. The main

TABLE 11-3 ReadSeq Servers Available on Internet

There are many others

Source	URL
Baylor College of Medicine	▸http://searchlauncher.bcm.tmc.edu/seq-util/readseq.html
Center for Information Technology, National Institutes of Health	▸http://bimas.dcrt.nih.gov/molbio/readseq/
Pasteur Institute	▸http://bioweb.pasteur.fr/seqanal/interfaces/readseq-simple.html
European Bioinformatics Institute	▸http://www.ebi.ac.uk/readseq/

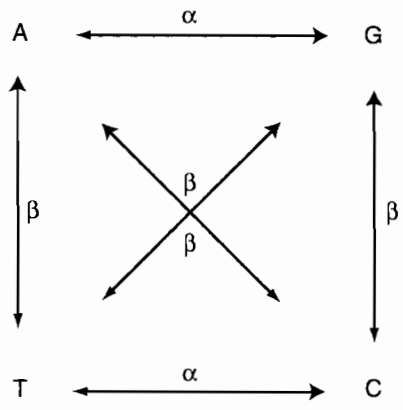


FIGURE 11.14. *The Jukes–Cantor model of evolution corrects for superimposed changes in an alignment. The model assumes that each nucleotide residue is equally likely to change to any of the other three residues and that the four bases are present in equal proportions. The rate of transitions (α) equals the rate of transversions (β). In the Kimura two-parameter model, $\alpha \neq \beta$. Typically, transversions are given more weight.*

reason is that character *differences* are not the same as *distances*: The differences between two sequences are easy to measure, but the genetic distance involves many mutations that cannot be observed directly. As shown in Figure 11.11, there are many kinds of mutations that occur but are not detected in an estimate of divergence based on counting differences. Jukes and Cantor proposed a corrective formula in 1969 (p. 100):

$$D = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right) \tag{11.6}$$

The Jukes–Cantor one-parameter model describes the probability that each nucleotide will mutate to another (Fig. 11.14). It makes the simplifying assumption that each residue is equally likely to change to any of the other three residues and that the four bases are present in equal proportions. Thus, this model assumes that the rate of transitions equals the rate of transversions. The corrections are minimal for very closely related sequences but can be substantial for more distantly related sequences. Beyond about 70% differences, the corrected distances are difficult to estimate.

Often, the transition rate is greater than the transversion rate. A variety of models have been developed that are more sophisticated than Jukes–Cantor. The Kimura two-parameter model adjusts the transition and transversion ratios by giving more weight to transversions to account for their likelihood of causing nonsynonymous changes in protein-coding regions. In any region of DNA (including noncoding sequence), the transition/transversion ratio corrects for the biophysical threshold for creating a purine-purine or pyrimidine-pyrimidine pair in the double helix. Other adjustments are to rate variations among sites (e.g., viruses or immunoglobulins often display hypervariable regions of mutation) or rate variation at different codon positions.

Making Trees Using UPGMA Distance-Based Method

There are many software programs available for making phylogenetic trees. We will focus on the Phylogeny Analysis Using Parsimony (PAUP) program. This is one of the most versatile, comprehensive, and popular programs available to make trees. Data can be imported from a variety of multiple sequence alignment formats, and analyses are possible using parsimony, distance, or other criteria (Fig. 11.15).

We can make a distance-based tree in PAUP simply by selecting the distance criterion from the analysis menu (Fig. 11.15*d*), then choosing UPGMA. A dialog box

An inherent limitation of distance-based tree-building methods is that character data are discarded. One thing this means is that when two aligned sequences differ at a particular residue, distance-based methods typically cannot interpret whether further mutations have occurred at that site. Character-based approaches such as parsimony also discard data and make simplifying assumptions (see below).

As an example of how to use Equation 11.6, consider an alignment where 3 nucleotides out of 60 aligned residues differ. The normalized Hamming distance is $\frac{3}{60} = 0.05$. The Jukes–Cantor correction $d = -\frac{3}{4}\ln[1 - (4 \times 0.05/3)] = 0.052$. In this case, applying the correction causes only a small effect. When $\frac{30}{60}$ nucleotides differ, the Jukes–Cantor correction is $-\frac{3}{4}\ln(1 - [4 \times 0.5/3]) = 0.82$, a far more substantial adjustment.

PAUP was written by David Swofford, currently of Florida State University. It is available from Sinauer Associates (▸http://www.sinauer.com) for most computer platforms. The program comes with helpful tutorials. For more information see ▸http://paup.csit.fsu.edu/. In this chapter we will focus on the Macintosh interface for PAUP.

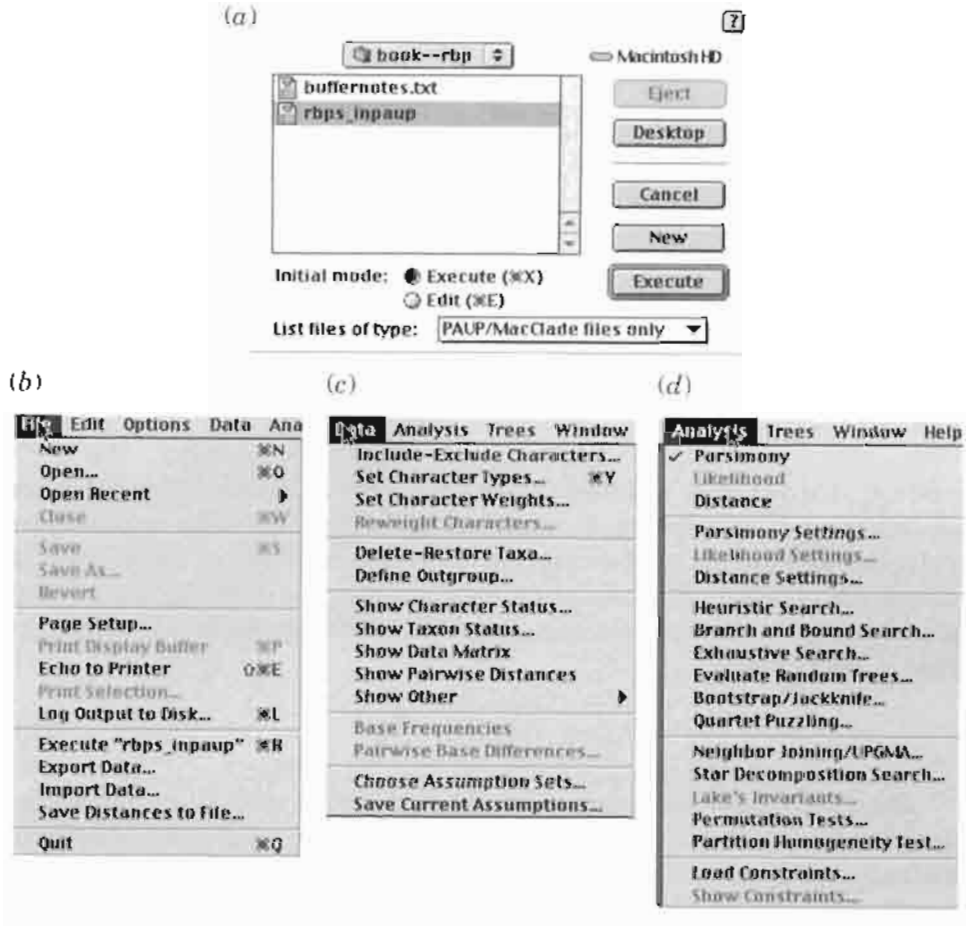


FIGURE 11.15. *The PAUP program, written by David Swofford of the Smithsonian Institution, provides comprehensive tools for phylogeny. Multiple sequence alignments can be imported from sources such as the Genetics Computer Group (GCG) multiple sequence format (MSF) files or ClustalW (Chapter 10), through* (a) *a dialog box from* (b) *the main File menu.* (c) *Once imported and executed, the data can be analyzed a variety of ways. For example, through "Include-Exclude Characters" it is easy to perform phylogenetic analyses on selected regions of a protein (or nucleic acid) alignment.* (d) *The analysis tools permit either parsimony or distance-based approaches.*

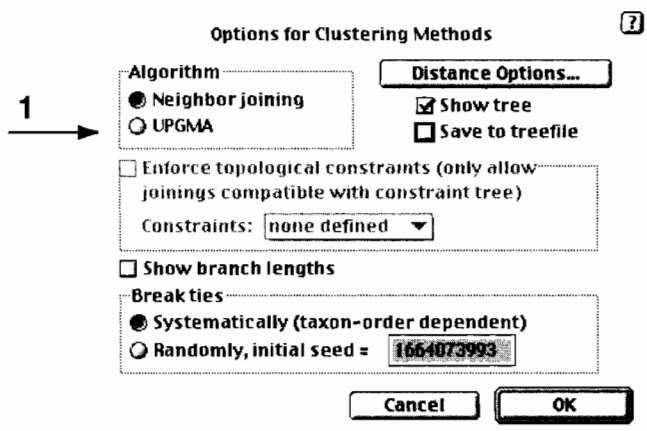
We described the use of a distance matrix to create a guide tree in Chapter 10.

appears to allow you to choose either an UPGMA or neighbor-joining algorithm (Fig. 11.16*a*). The UPGMA is a simple tree-making algorithm that works by clustering the sequences based on a distance matrix. We introduce UPGMA here because the tree-building process is relatively intuitive. However, the algorithm most experts employ to build distance-based trees is neighbor-joining (described below). As the clusters grow, a tree is assembled. A tree of 13 RBPs using UPGMA is shown in Figure 11.16*b*. As we would expect, the four fish RBP sequences are clustered away from the others, forming a distinct clade. The two most closely related proteins, rat and mouse RBP, share 99.5% amino acid identity. They are clustered most closely together. Note that the mammalian RBPs form a distinct clade as well, separate from the avian and amphibian proteins.

The UPGMA algorithm is relatively simple. Consider five sequences whose distances can be represented as points in a plane (Fig. 11.17*a*). Some protein sequences, such as 1 and 2, are closely similar, while others (such as 1 and 3) are far less related. UPGMA clusters the sequences as follows (adapted from Durbin et al., 1998, pp. 166 ff):

1. First, each sequence i is assigned to its own cluster C_i. In our example, there are five isolated clusters. Each is placed as an OTU (or leaf) at height zero on the tree. For our example, we will place the numbers 1–5 on the x axis as the terminal nodes of a tree.

(*a*)



(*b*)

```
Processing of file "rbps_inpaup" begins...

Data read in 'protein' format

Data matrix has 13 taxa, 219 characters
Valid character-state symbols: ACDEFGHIKLMNPQRSTVWY*
Missing data identified by '-'
"Equate" macros in effect:
   B,b ==> {DN}
   Z,z ==> {EQ}

Processing of file "rbps_inpaup" completed.

Optimality criterion set to distance.

UPGMA search settings:
  Ties (if encountered) will be broken systematically
  Distance measure = mean character difference
  (Tree is rooted)

   Tree found by UPGMA method stored in tree buffer
   Time used = 0.00 sec

UPGMA tree:
```

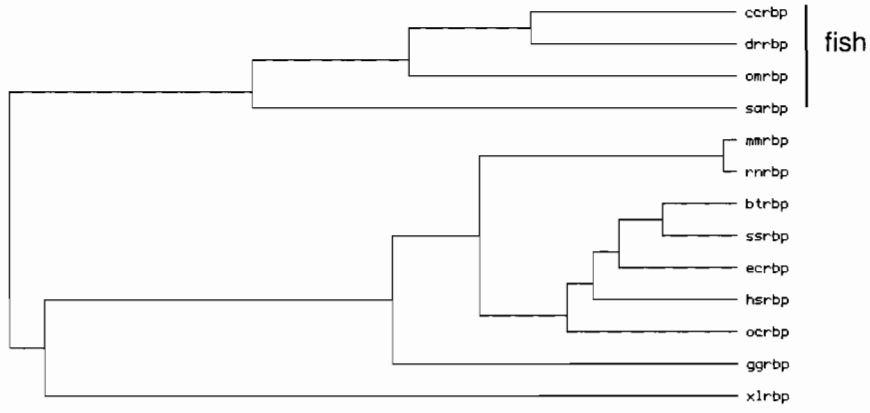


FIGURE 11.16. *One of the simplest algorithms for making a tree is the distance-based UPGMA.* (a) *This is implemented in PAUP by setting the Analysis menu to "Distance" (see Fig. 11.15d), then choosing UPGMA.* (b) *The tree that is generated represents the 13 RBPs. Note that the fish RBPs are grouped together, as expected. The rat and mouse sequences are most closely related on the tree; these two proteins share 99.5% identity.*

2. The two clusters C_i, C_j are determined for which the distance between them, d_{ij}, is minimal. If there are several equidistant minimal pairs, one is picked randomly. In our example, protein sequences 1 and 2 are first clustered (Fig. 11.17*b*). Note that we can immediately define an internal node. We number it 6 because numbers 1–5 are reserved for the external nodes.
3. A new cluster k is defined by $C_k = C_i \cup C_j$.

   A node k is defined with daughter nodes i and j, and it is placed at height $\frac{1}{2}d_{ij}$. To the current clusters k is added and i and j removed.

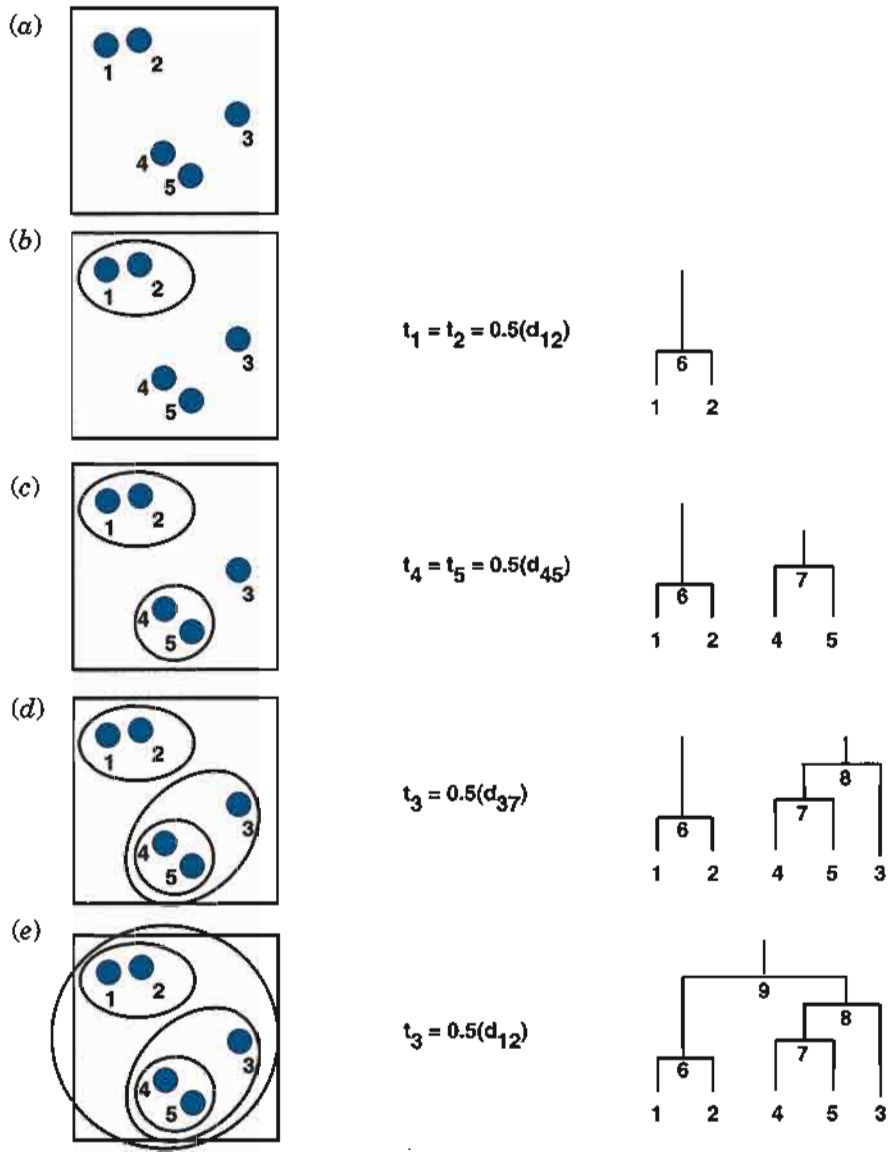


FIGURE 11.17. Explanation of the UPGMA. This is a simple, fast algorithm for making trees. It is based on clustering sequences. (a) *Each sequence is assigned to its own cluster.* (b) *The taxa with the closest distance (sequences 1 and 2) are identified and connected. This allows us to name an internal node [right, node 6, in* (b)*].* (c) *The next closest sequences are 4 and 5, then* (d) *the newly formed group 45 with sequence 3. Finally,* (c) *all sequences are connected in a tree.*

4. These iterations are continued. As each new cluster is defined, a new distance matrix is calculated by replacing the rows for C_i and C_j with one for C_k, and then calculating distances by averaging distance values for C_i and C_j. In our example, $d_{3(1+2)} = (d_{31} + d_{32})/2$.
5. When only two clusters i, j remain, the root is placed at height $\frac{1}{2}d_{ij}$.

An UPGMA tree is always rooted. A critical assumption of the UPGMA approach is that the molecular clock is constant for the sequences in the tree. If this assumption is true, branch lengths can be used to estimate the dates of divergence. If it is violated and there are unequal substitution rates along different branches of the tree, the method can produce an incorrect tree.

The UPGMA method is presented here largely because it is instructive as a distance method. Also, it is commonly used for microarray data analysis (Box 11.3). David Hillis and others have demonstrated that for molecular sequence data it is significantly less accurate than methods such as neighbor-joining.

BOX 11-3
Trees from Microarray Data Versus Phylogenetic Trees

In trees generated from microarray data (such as shown in Fig. 7.10 or 7.12), elements that are very similar to each other are joined by short branches. Longer branches are associated with dissimilarity. Phylogenetic trees are constructed using UPGMA or similar algorithms, and again more distantly related objects have longer branches. Overall, the trees may look the same. The strength of distance methods is that almost any matrix can be used to construct a tree. So these algorithms are applied to many types of data.

Making Trees by Distance-Based Methods: Neighbor Joining (NJ)

The NJ method is used for building trees by distance methods (Saitou and Nei, 1987). We begin by generating a full tree with a starlike structure (Fig. 11.18*a*). Pairwise comparisons are then made to identify the two most closely related sequences. These are neighbors, which are defined as OTUs connected through a single node (Fig. 11.18*b*). This process is continued until the topology of the tree is completed: neighbors 1 and 2, which form a clade, are next matched to OTU 3.

Any pair of OTUs can form the first cluster. There are $\frac{1}{2}N(N-1)$ possible neighbor pairs.

The NJ algorithm minimizes the sum of branch lengths at each stage of clustering OTUs. Thus, it is sometimes called a "minimum-evolution" approach. It produces both a tree topology and an estimate of the branch lengths. An example of an NJ tree for 13 RBPs is shown in Figure 11.19. This algorithm is especially useful when studying large numbers of taxa.

The HIV Sequence Database at the Los Alamos National Laboratory offers Neighbor TreeMaker, an online program for making phylogenetic trees. It accepts a multiple sequence alignment as input, then uses the PHYLIP program's Dnadist (a distance matrix program) and Neighbor (a treefile generator), and then displays a tree. Visit ▸http://hiv-web.lanl.gov/content/hiv-db/CONTAM/TreeMaker/TreeMaker.html.

Making Trees by Character-Based Methods

Character-based methods analyze candidate trees based on relationships inferred directly from the sequence alignment. These approaches are distinct from distance-based methods, because they do not involve an intermediary summary of the sequence data in the form of a distance matrix. Nonetheless they do depend on distance metrics.

Maximum Parsimony (MP)

The main idea behind MP is that the best tree is the tree with the shortest branch lengths possible (Czelusniak et al., 1990). According to MP theory, having fewer changes to account for the way a group of sequences evolved is preferable to more complicated explanations of molecular evolution. Thus we seek the most parsimonius explanations for the observed data. The steps are as follows:

- Identify informative sites. If a site is constant (e.g., Fig. 11.13, arrows 2, 4, and 5), then it is not informative (see below). Position 43 on that alignment (two columns to the right of arrow 3) is also not parsimony-informative, although its characters are not constant. 11 taxa have a serine, one has an alanine, and one has an aspartate. Parsimony-informative characters must have at least two states that occur in at least two taxa.
- Construct trees. Every tree is assigned a cost, and the tree with the lowest cost is sought. When a reasonable number of taxa are evaluated, such as about a dozen or fewer, all possible trees are evaluated and the one with the shortest branch length is chosen. When necessary, a heuristic search is performed to

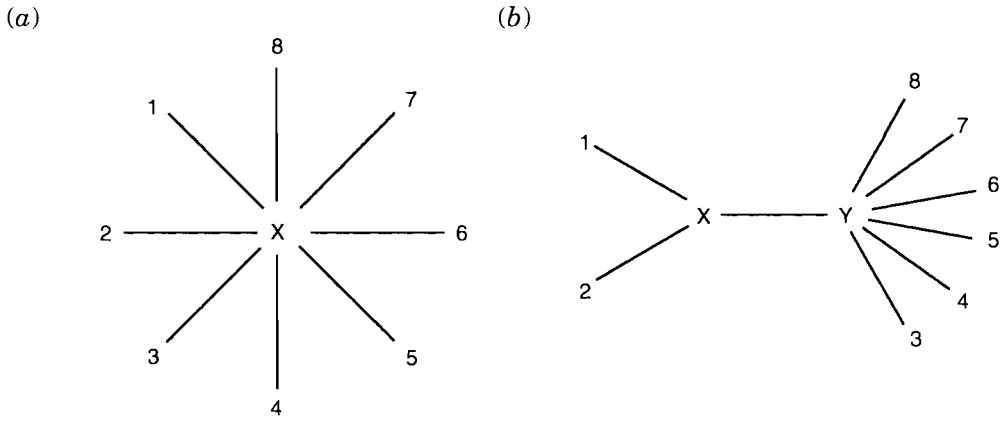


FIGURE 11.18. *The NJ method is a distance-based algorithm.* (a) *The OTUs are first clustered in a starlike tree. "Neighbors" are defined as OTUs that are connected by a single, interior node in an unrooted, bifurcating tree.* (b) *The two closest OTUs are identified, such as OTUs 1 and 2. These neighbors are connected to the other OTUs via the internal branch XY. The OTUs that are selected as neighbors in* (b) *are chosen as the ones that yield the smallest sum of branch lengths. This process is repeated until the entire tree is generated. Adapted from Saitou and Nei (1987). Used with permission.*

reduce the complexity of the search by ignoring large families of trees that are unlikely to contain the shortest tree.

- Count the number of changes and select the shortest tree (or trees).

As an example of how MP works, consider four aligned nucleotide sequences (Fig. 11.20). Three possible trees describe these sequences; each tree has hypothetical sequences assigned to ancestral nodes. One of the trees (Fig. 11.20, lower left) requires the fewest changes to explain how the observed sequences evolved from a hypothetical common ancestor. In this example, each site is treated independently. Both weighted and unweighted parsimony algorithms have been developed; weighting adds costs for particular nucleotide or amino acid substitutions. For example, nucleotide transversions are more penalized than transitions.

In PAUP, you can set the tree-making criterion to parsimony (see Fig. 11.15*d*). It would be ideal to perform an exhaustive search of all possible trees to find the one with the shortest total branch lengths. In practice, this is not possible for more than about 10 or 12 taxa, so it is necessary to perform a heuristic search (Fig. 11.21). Both heuristic and exhaustive searches often result in the identification of several trees having the same minimal value for total branch length of the tree. Trees can be visualized as a phylogram (Figs. 11.22*a*, *b*) or a cladogram (Figs. 11.22*c*, *d*). In the phylogram, branch lengths are proportional. Thus taxa that appear close to each other in the tree are more evolutionarily related. In contrast, in the cladograms the taxa are spread out. This has the advantage of allowing all the branch length values to be easily visible, and the labels are neatly arranged (Fig. 11.22*d*), although much of the visual impact is lost.

An artifact called long-branch attraction sometimes occurs in parsimony studies. In a phylogenetic reconstruction of protein or DNA sequences, a branch length indicates the number of substitutions that occur between two taxa. Parsimony algorithms assume that all taxa evolve at the same rate and that all characters contribute the same amount of information. Long-branch attraction is a phenomenon in which rapidly evolving taxa are placed together on a tree, not because they are closely related, but artifactually because they both have many mutations. Consider the true tree in Figure 11.23, in which taxon 2 represents a DNA or protein that

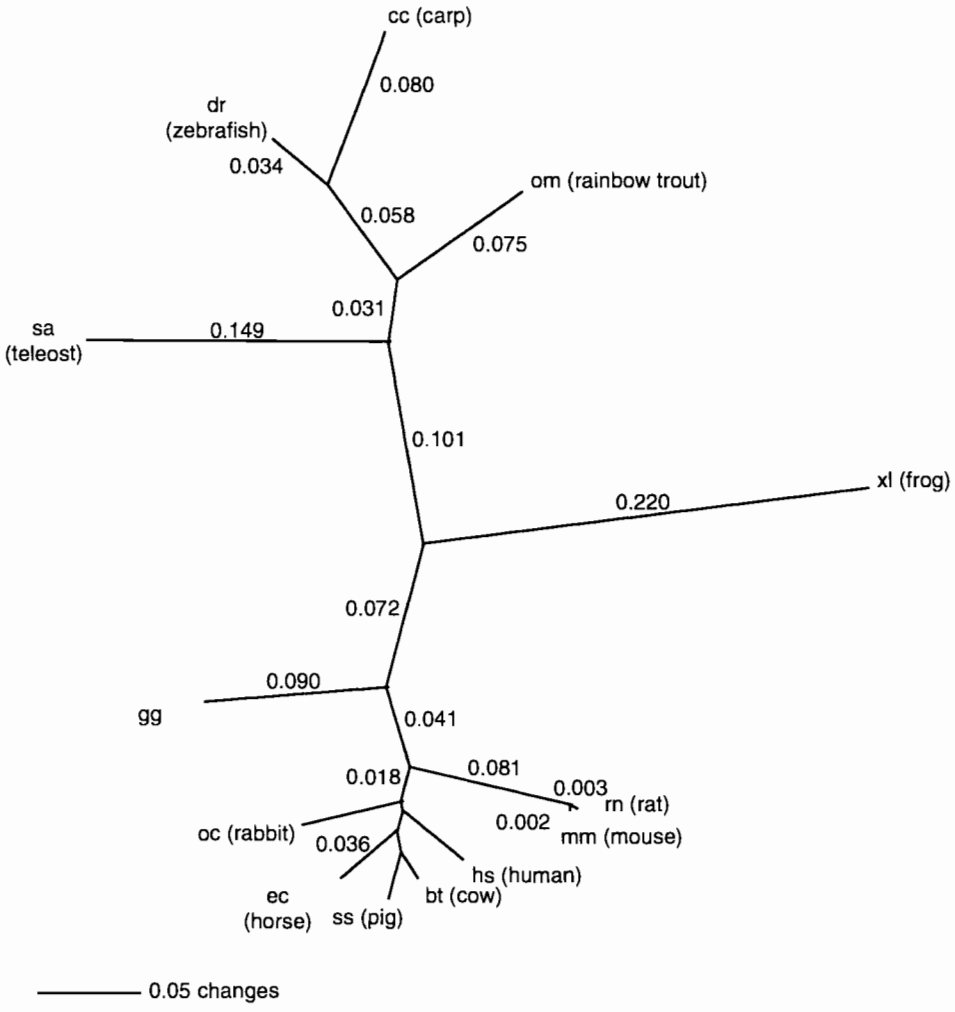


FIGURE 11.19. *Phylogenetic analysis of 13 RBPs by NJ using a distance criterion. The analysis was done in PAUP, and the tree is in the form of an unrooted phylogram. Several of the distance values are omitted for clarity. See Figure 3.2 for abbreviations and accession numbers.*

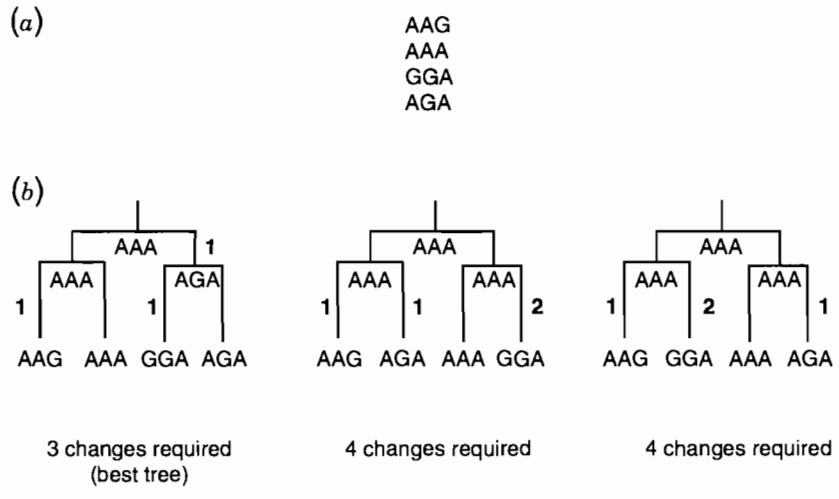


FIGURE 11.20. *Principle of MP.* (a) *Consider the four aligned taxa (i.e., sequences), each three nucleotides in length. Three trees that describe how these sequences could have evolved are shown.* (b) *The tree requiring the fewest changes (at left) is selected by MP algorithms. From Durbin et al. (1998, p. 173). Used with permission.*

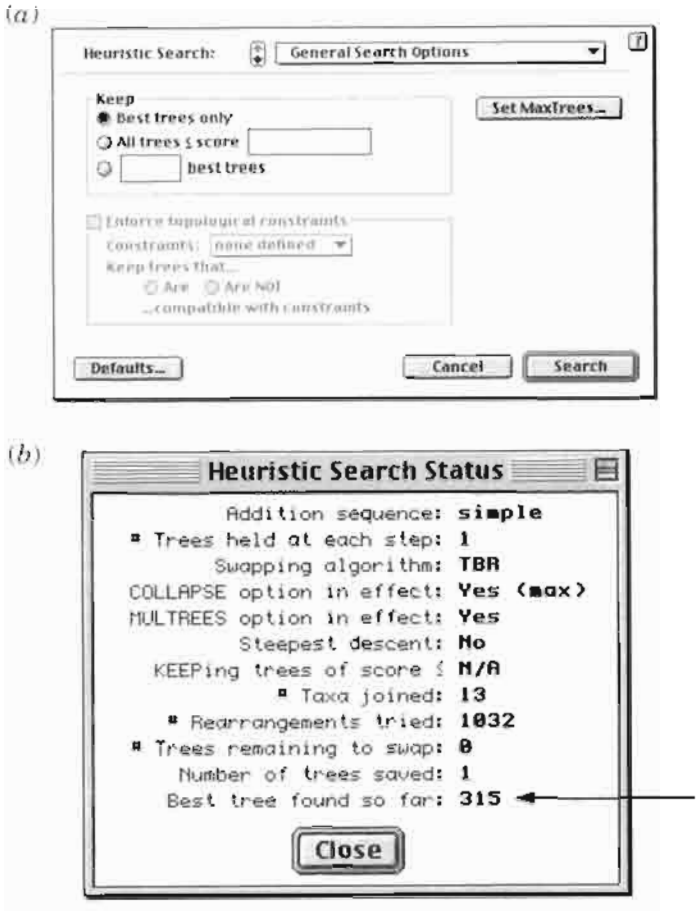


FIGURE 11.21. (a) *A heuristic search can be performed in PAUP.* (b) *The total branch length of the best tree is reported (arrow). Using the parsimony criterion, it is common to find multiple trees sharing the same lowest score. A consensus tree of these optimal trees often includes multifurcations.*

changes rapidly relative to taxa 1 and 3. The outgroup is (by definition) more distantly related than taxa 1, 2, and 3 are to each other. An MP algorithm may generate an inferred tree (Fig. 11.23) in which taxon 2 is "attracted" toward another long branch (the outgroup) because these two taxa have a large number of substitutions. Anytime two long branches are present, they may be attracted.

Maximum Likelihood (ML)

Maximum parsimony methods can sometimes fail, especially when there are large amounts of evolutionary change in different branches of a tree. An alternative character-based approach is ML (Felsenstein, 1981). ML is among the most computationally intensive, but most flexible methods available. Maximum-likelihood algorithms search for a tree that has the highest probability of producing the observed data. A likelihood is calculated for each residue in an alignment, based upon some model of the substitution process. In the simplest case, this model could be that all nucleotide or amino acid changes are equally probable. Maximum-likelihood approaches usually include parameters to account for different substitution rates. For example, the TREE-PUZZLE program (Strimmer and von Haeselser, 1996; Schmidt et al., 2002) accounts for rate heterogeneity across sites.

The TREE-PUZZLE program of Korbinian Strimmer and colleagues is available at ▸ http://www.tree-puzzle.de/. You can perform ML using DNAML (PHYLIP) and PAUP. The MODELTEST program of Posada and Crandall (1998) uses log likelihood scores to select from 56 models of DNA evolution that best fits the data. See ▸ http://inbio.byu.edu/Faculty/kac/crandall-lab/modeltest.htm

Stage 4: Evaluating Trees Using Randomizing Tests and Bootstrapping

After you have constructed a phylogenetic tree, how can you assess its accuracy? The main criteria by which accuracy may be assessed are consistency, efficiency,

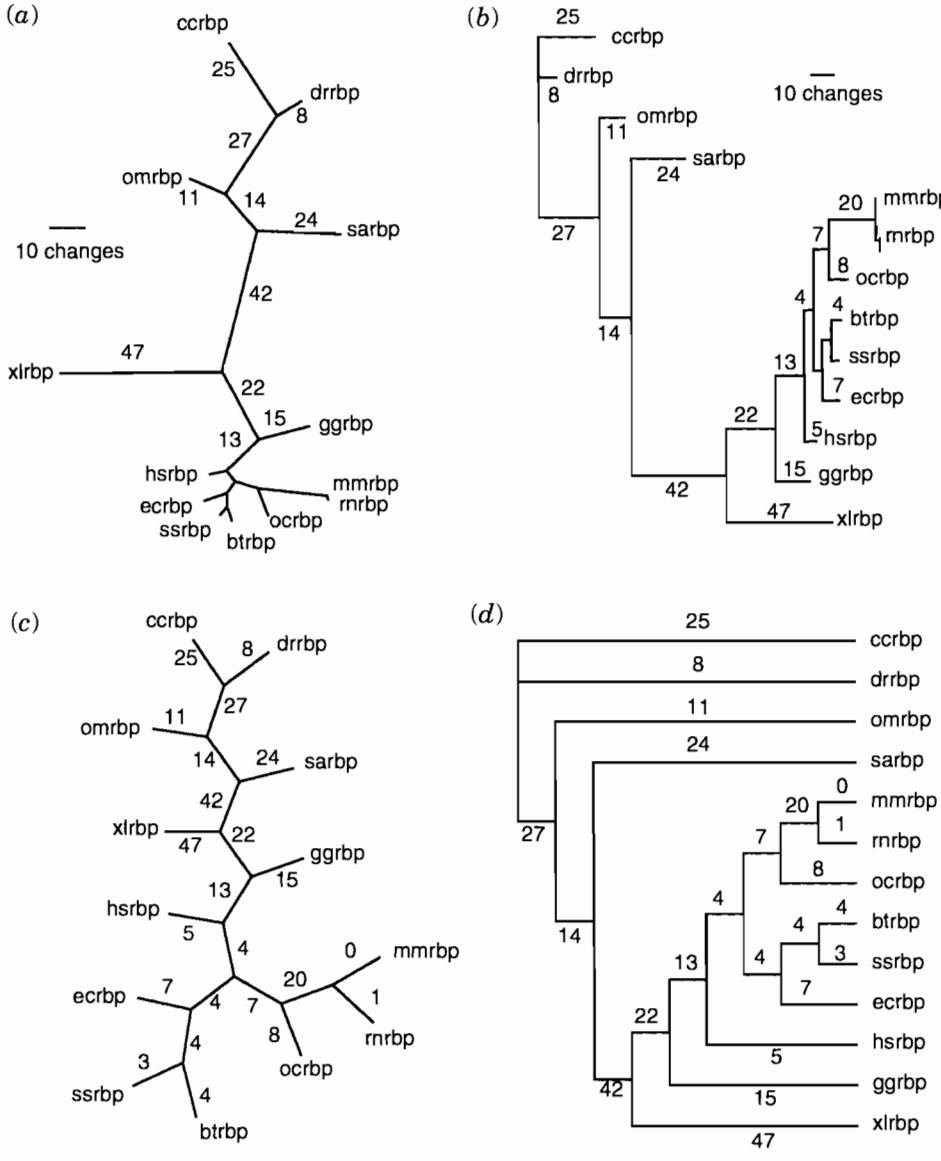


FIGURE 11.22. *Phylogenetic analysis of 13 RBPs using a criterion of parsimony. The same data are depicted in the form of* (a) *a phylogram,* (b) *a rectangular phylogram,* (c) *a cladogram, and* (d) *a rectangular cladogram. Each graph has different advantages. Values are proportional to branch length in* (a) *and* (b) *(note scale bars); not all values are shown in these two plots. In* (c) *and* (d) *values are not proportional to branch lengths, and the values are easier to see than in* (a) *and* (b)*, where closely related sequences appear clustered together.*

and robustness (Hillis, 1995; Hillis and Huelsenbeck, 1992). One may study the accuracy of a tree-building approach or the accuracy of a particular tree. The most common approach is bootstrap analysis (Hillis and Bull, 1993). Bootstrapping is not a technique to assess the accuracy of a tree. Instead, it describes the robustness of the tree topology: Given a particular branching order, how consistently does a

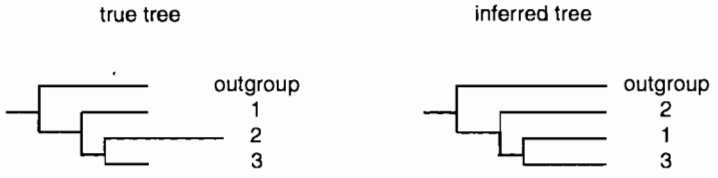
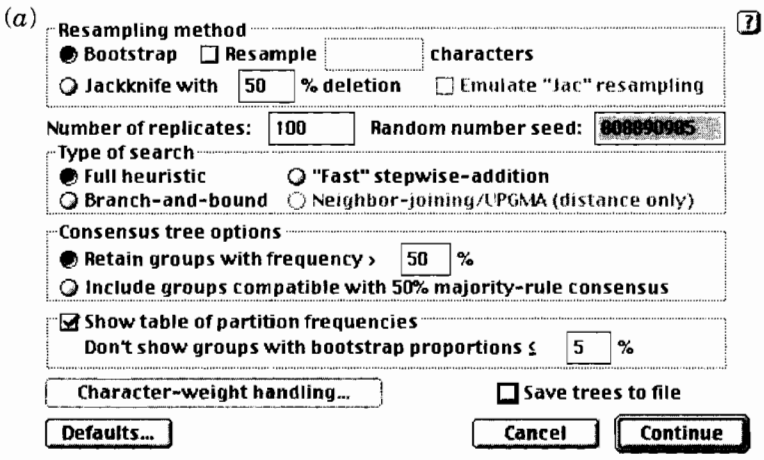


FIGURE 11.23. *Long-branch-chain attraction. The true tree includes a taxon (labeled 2) that evolves more quickly than the other taxa. It shares a common ancestor with taxon 3. However, the inferred tree taxon 2 is placed separately from the other taxa because it is attracted by the long branch of the outgroup. Adapted from Philippe and Laurent (1998).*



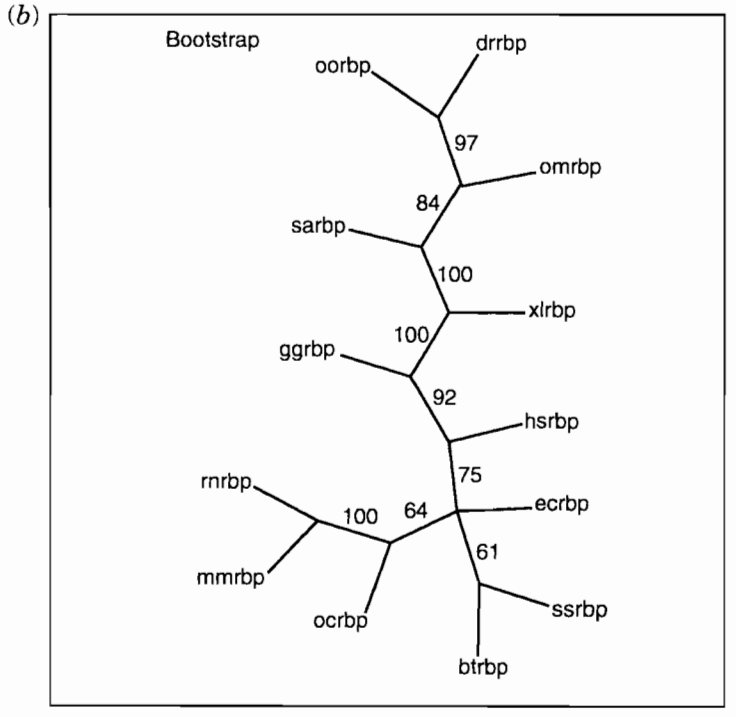


FIGURE 11.24. (a) *Bootstrapping is accomplished in PAUP. Jackknife is a complementary resampling method that often yields similar results. Typically, 100–1000 bootstrap replicates are executed.* (b) *The output includes a tree in which the values are the percent of instances in which bootstrap trees support each clade in the inferred tree. For example, in 100% of the bootstrap trials, mouse and rat RBP* (mmrbp *and* rnrbp) *were supported as being in a clade. However, the clade containing mouse, rat, and rabbit* (ocrbp) *proteins was supported in only 64% of the bootstrap replicates. This means that in 36% of the bootstrap trees, either a protein such as equine RBP* (ecrbp) *joined that group of three proteins or else one of the three proteins (e.g.,* ocrbp) *joined a different clade. Thus, the bootstrap can provide a measure of how well supported an inferred tree topology is upon repeated samplings of the data set.*

Accuracy refers to the degree to which a tree approximates the true tree.

Parametric bootstrapping refers to repeated random sampling without replacement from the original sample. The result is a shorter alignment, relying on less of the data. It is not used as often as nonparametric bootstrapping.

tree-building algorithm find that branching order using a randomly-permuted version of the original data set?

Nonparametric bootstrapping is performed as follows. A multiple sequence alignment is used as the input data to generate a tree using some tree-building method. The program then makes an artificial data set of the same size as the original data set by randomly picking columns from the multiple sequence alignment. This is usually performed with replacement, meaning that any individual column may appear multiple times (or not at all). A tree is generated from the randomized data set. A large number of bootstrap replicates are then generated; that is, between 100 and 1000 new trees are made by this process. The bootstrap trees are compared to the original, inferred trees. The information you get from bootstrapping is the frequency with which each clade in the original tree is observed.

An example of the bootstrap procedure using PAUP is shown in Figure 11.24. The percentage of times that a given clade is supported in the original tree is

provided based on how often the bootstraps supported the original tree topology (Fig. 11.24*b*). Bootstrap values above 70% are generally considered to provide support for the clade designations. Hillis and Bull (1993) have estimated that such values provide statistical significance at the $p < 0.05$ level. This approach measures the effect of random weighting of characters in the original data matrix, giving insight into how strongly the phylogenetic signal that produces a tree is distributed through the multiple sequence alignment.

PERSPECTIVE

Molecular phylogeny is a fundamental tool for understanding the evolution and relationships of protein (and nucleic acid) sequences. The main output of this analysis is a phylogenetic tree, which is a graphical representation of a multiple sequence alignment. The recent rapid growth of DNA and protein sequence data, along with the visual impact of phylogenetic trees, has made phylogeny increasingly important and widely applied. We will show examples of trees in Chapters 12–16.

PITFALLS

The quality of a phylogenetic tree based on molecular sequence data depends upon the quality of the sequence data and the multiple sequence alignment. It is also necessary to choose the appropriate model for the phylogeny. Finally, your understanding of the output of the phylogenetic analysis is critical. Each of the methods used to reconstruct phylogenetic trees involves many assumptions and suffers from potential weaknesses.

WEB RESOURCES

The best starting point for phylogeny resources on the World Wide Web is the site of Joe Felsenstein (▸http://evolution.genetics.washington.edu/phylip/software.html). About 200 links are listed, organized by categories such as phylogenetic methods, computer platforms, and assorted types of data.

DISCUSSION QUESTION

[11-1] Are there gene (or protein) families for which you expect distance-based tree-building methods to give substantially different results than character-based methods?

PROBLEM

[11-1] **Alignment Input**

1. Go to Conserved Domain Database (CCD) at NCBI (▸http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml).
2. At the bottom right of the homepage for CDD you will find the option to "Find CDs" by keyword; type in "lipocalins." Click on "Find CDs." The search result will be "pfam00061." Click on the family.
3. The new window will present you with the brief introduction of the lipocalin protein family, including the representative consensus sequence created from 59 aligned sequences. (Note that there are more than

59 lipocalins, but 59 is the number represented in the PFAM entry.)

4. Select mFasta then click on "View Alignment."
5. Copy the FASTA alignment into a simple text file. Change names of the sequences from *gi number* into a name having up to nine characters in the first row of the alignment. It is helpful to change the filenames as follows: convert "gi|809398" to "btrbp" (*Bos taurus* retinol binding protein). PAUP does not usually tolerate "/" and numbers. Also change the consensus sequence by deleting the text up to the beginning of the protein sequence and name it with a nine-character name ("consensus").

Convert to Readable Format (via ReadSeq)

6. Open a new window in your browser and go to a ReadSeq page such as ▸http://searchlauncher.bcm.tmc.edu/seq-util/readseq.html. You can find this URL easily by entering the query "readseq" into a search engine.
7. Paste your FASTA sequences into the provided window. Choose the PAUP/NEXUS output format. Double click on "Perform Conversion."
8. Take the output alignment, and copy it into your computer's clipboard. Alternatively copy it into simple text or a notepad file. Start copying with the #NEXUS and end with the symbols "end;".

Import into PAUP

9. Open the PAUP program.
10. Under the File heading, choose "New."
11. Paste in your alignment.
12. Under the File menu, choose "Execute."
13. If you get some error message you may have to try several alternatives such as
    (i) Look for the line "format datatype=protein interleave missing=-;" and change the hyphen (-) in this line to a period (.)
    (ii) Under the Edit menu choose "Find." Replace all instances of the symbol / with the underscore (-).
14. When the Execute command is successful, you will be notified that "Data matrix has 10 taxa (i.e., 10 proteins) and 163 characters (amino acids)."

Tree Analysis Based on Maximum Parsimony (MP)
(a) Heuristic Search

15. Under the Analysis heading, select "Parsimony."
16. Under the Analysis menu, choose "Heuristic search." Click "Search" and use the default setting to find the tree.
17. Once the search finishes, the small window with "Heuristic Search Status" will have a clickable close icon. The program describes the number of tree rearrangements that were tried and the scores for the best tree(s).
18. View the tree. Go under the Trees menu and choose "Print Tree." You will have the option to see an unrooted phylogram or cladogram among several options. In phylogram displays, the branch lengths are proportional to amino acid changes, and the tree is accompanied by a scale bar. On the other hand, branch lengths are not proportional to amino acid changes in the cladogram. A cladogram portrays evolutionary relationships within species and populations. In addition, if you have more than one tree found to have the same best score, you will have the option to view the particular tree individually or to display all trees at once.
19. How many amino acid changes occur in the shortest branch and longest branch in your tree? Which OTUs (taxa) are connected by these branches?
20. If you have more than one tree, you can choose a consensus tree. Under the Trees menu, select "Compute consensus."

(b) Evaluation of Trees

(i) The principle of the random tree test is to compare the score of the found tree to the score distribution of X randomly generated trees starting with your alignment.

21. Evaluate the tree by frequency distribution of lengths of 100, 1000, and 10,000 random trees.
22. Go under the Analysis icon and use the "Evaluate random tree." option. Change the number of random trees and see how the mean and standard deviation change.
23. Is the score of your tree found by a heuristic search significantly better than the score distribution from randomly generated trees?
24. Perform the heuristic search for the most parsimonious tree again and check the score. How good is your score this time? Since we are working with the protein alignment of 10 taxa (OTUs), we can perform the exhaustive search for the tree with maximum parsimony. Go under the Analysis icon and choose "Exhaustive search." What score did you get and how does it compare to the score(s) obtained using the heuristic search strategy?

(ii) The bootstrap test is another type of resampling test. The principle is to randomly sample the individual columns of aligned amino acid sequence data from the original alignment. The newly

generated data sets will maintain the identical size of the original alignment. The bootstrap describes the percent of instances in which a particular clade designation is supported.

25. Perform the bootstrap test. Go under the Analysis icon and choose "Bootstrap/Jacknife." Then change the number of sampling. Try 1000 and 10,000 replicate samplings with replacement.
26. After each test you can view the tree → Trees → Print Bootstrap Consensus → Preview. (The plot type can be changed to "unrooted.")
27. Analyze the tree with 10,000 replicate samplings. Based on your bootstrap values, how many strongly-supported clades (bootstrap value >70%) are present in your tree and what taxa do they comprise?
28. Can you determine if members of the particular clade are paralogs or orthologs? What kind of information do you need to make this inference?

Change Input Files

29. Go back into the CDD database and retrieve the lipocalin "pfam00061" family (points 1-3 on page 389).
30. Instead of choosing "10 most diverse sequences" as is set by default (see point 4), choose "top listed sequences" in the option of sequences in the output alignment. (Alternatively, e.g., you can increase the input of you sequences up to 25.)
31. Repeat the above exercise and observe the differences.
32. You can also customize the input of the sequences. Choose the option "Selected sequences" and check the chosen sequences listed below this menu (see point 4).
33. *Analysis of trees based on distance method.* This method is based on comparing the number of pairwise differences in sequences and using the computed distances between the sequences to construct a tree. Unfortunately, some of these mutations (especially if you are constructing a DNA rather than a protein tree) can become overlooked if a mutation occurs following another mutation back to the original character.

    Under the Analysis icon choose the "Distance" option. Then perform the heuristic search for the tree. View the tree and see if the taxa (OTUs) separated as they did in a tree based on maximum parsimony. In addition, perform an evaluation of your tree with randomness testing and bootstrap as you did with the tree above.
34. *Analysis of trees based on maximum likelihood method.* This is a character-based tree-building method (as is maximum parsimony). In this case, the trees are evaluated based on the likelihood of producing the observed data. The PAUP program will let you to construct the tree only with aligned DNA sequences.

    One way to obtain aligned DNA sequences is to retrieve the alignment from the PopSet database. (Go to the main NCBI site and choose Entrez → click PopSet → browse the Popset for your favorite alignment). Paste the PHYLIP formatted alignment from PopSet into ReadSeq and proceed as above.

    A tree based on the maximum-likelihood method from protein alignment can be created with program Puzzle (▸http://www.tree-puzzle.de).

SELF-TEST QUIZ

[11-1] According to the molecular clock hypothesis:
- (a) All proteins evolve at the same, constant rate.
- (b) All proteins evolve at a rate that matches the fossil record.
- (c) For every given protein, the rate of molecular evolution gradually slows down like a clock that runs down.
- (d) For every given protein, the rate of molecular evolution is approximately constant in all evolutionary lineages.

[11-2] The two main features of any phylogenetic tree are:
- (a) The clades and the nodes
- (b) The topology and the branch lengths
- (c) The clades and the root
- (d) The alignment and the bootstrap

[11-3] Which one of the following is a character-based phylogenetic algorithm?
- (a) Neighbor joining
- (b) Kimura
- (c) Maximum likelihood
- (d) PAUP

[11-4] There are two basic ways to make a phylogenetic tree: distance based and character based. The fundamental difference between them is:
- (a) Distance-based methods include an intermediary data matrix used to define the branching order while character-based methods do not.
- (b) Distance-based methods are only used for DNA data while character-based methods are used for DNA or protein data.
- (c) Distance-based methods use parsimony while character-based methods do not.
- (d) Distance-based methods have branches that are proportional to time while character-based methods do not.

[11-5] An example of an operational taxonomic unit (OTU) is:
(a) Multiple sequence alignment
(b) Protein sequence
(c) Clade
(d) Node

[11-6] For a given pair of OTUs, which of the following is true?
(a) The corrected genetic distance is greater than or equal to the proportion of substitutions.
(b) The proportion of substitutions is greater than or equal to the corrected genetic distance.

[11-7] Transitions are almost always weighted more heavily than transversions.
(a) True
(b) False

[11-8] One of the most common errors in making and analyzing a phylogenetic tree is:
(a) Using a bad multiple sequence alignment as input
(b) Trying to infer the evolutionary relationships of genes (or proteins) in the tree
(c) Trying to infer the age at which genes (or proteins) diverged from each other
(d) Assuming that clades are monophyletic

[11-9] Clustal X can be used to generate neighbor-joining trees with or without bootstrap values.
(a) True
(b) False

[11-10] You have 200 viral DNA sequences of 500 residues each, and you want to know if there are any pairs that are identical (or nearly identical). Which of the following is the most efficient method to use?
(a) BLAST
(b) Maximum-likelihod phylogenetic analysis
(c) Neighbor-joining phylogenetic analysis
(d) Popset

Suggested Reading

An excellent overview of evolution from a history-of-science perspective is provided by Mayr (1982). There are many superb textbooks on molecular evolution, including those by Graur and Li (2000) and Li (1997). A book by Maddison and Maddison (2000) describing the MacClade software package provides an extensive, clear introduction to phylogeny. Hall's book (2001) provides another excellent, practical introduction.

The National Center for Biotechnology Information offers an online primer (*Systematics and Molecular Phylogenetics*) that introduces molecular trees (▸http://www.ncbi.nlm.nih.gov/About/primer/phylo.html).

A thorough, detailed overview of phylogenetic inference is provided in a chapter by David Swofford (the author of PAUP) and colleagues (Swofford et al., 1996). A highly recommended overview of phylogenetics is by Thornton and DeSalle (2000). They describe phylogenetic principles (e.g., character-based and distance-based approaches), the assessment of orthology and paralogy, and the use of phylogeny in comparative genomics.

References

Anfinsen, C. B. The Molecular Basis of Evolution. John Wiley and Sons, New York, 1959.

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., and Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290,** 972–977 (2000).

Baxevanis, A. D., and Ouellette, B. F. *Bioinformatics,* 2nd ed. Wiley-Interscience, New York, 2001.

Cavalli-Sforza, L. L., and Edwards, A. W. F. Phylogenetic analysis: models and estimation procedures. *Amer. J. Hum. Genet.* **19,** 233–257 (1967).

Clote, P., and Backofen, R. *Computational Molecular Biology. An Introduction.* Wiley, New York, 2000.

Copley, R. R., Schultz, J., Ponting, C. P., and Bork, P. Protein families in multicellular organisms. *Curr. Opin. Struct. Biol.* **9,** 408–415 (1999).

Czelusniak, J., Goodman, M., Moncrief, N. D., and Kehoe, S. M. Maximum parsimony approach to construction of evolutionary trees from aligned homologous sequences. *Methods Enzymol.* **183,** 601–615 (1990).

Darwin, Charles. *The Origin of Species by Means of Natural Selection.* John Murray, London, 1859.

Dayhoff, M. O. *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Silver Spring, MD, 1978.

Dickerson, R. E. Sequence and structure homologies in bacterial and mammalian-type cytochromes. *J. Mol. Biol.* **57,** 1–15 (1971).

Dobzhansky, T. Nothing in biology makes sense except in the light of evolution. *Amer. Biol. Teacher* **35,** 125–129 (1973).

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. *Biological Sequence Analysis.* Cambridge University Press, Cambridge, 1998.

Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17,** 368–376 (1981).

Felsenstein, J. Phylogenies from molecular sequences: Inference and reliability. *Annu. Rev. Genet.* **22,** 521–565 (1988).

Felsenstein, J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. *Methods Enzymol.* **266,** 418–427 (1996).

Feng, D. F., and Doolittle, R. F. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.* **25,** 351–360 (1987).

Graur, D., and Li, W.-H. *Fundamentals of Molecular Evolution,* 2nd ed. Sinauer Associates, Sunderland, MA, 2000.

Haeckel, E. *The Riddle of the Universe.* Harper and Brothers, New York, 1900.

Hall, B. G. *Phylogenetic Trees Made Easy. A How-To for Molecular Biologists.* Sinauer Associates, Sunderland, MA, 2001.

Hein, J. Unified approach to alignment and phylogenies. *Methods Enzymol.* **183,** 626–645 (1990).

Hillis, D. M. Approaches for assessing phylogenetic accuracy. *Syst. Biol.* **44,** 3–16 (1995).

Hillis, D. M., and Bull, J. J. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biol.* **42,** 182–192 (1993).

Hillis, D. M., and Huelsenbeck, J. P. Signal, noise, and reliability in molecular phylogenetic analyses. *J. Hered.* **83,** 189–195 (1992).

Jukes, T. H., and Cantor, C. Evolution of protein molecules. In *Mammalian protein metabolism.* H. N. Munro and J. B. Allison, J. B. (eds.), Academic Press, New York, 1969 pp. 21–132.

Kimura, M. Evolutionary rate at the molecular level. *Nature* **217,** 624–626 (1968).

Kimura, M. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, 1983.

Li, W.-H. *Molecular Evolution.* Sinauer Associates, Sunderland, MA, 1997.

Maddison, D., and Maddison, W. *MacClade 4: Analysis of Phylogeny and Character Evolution.* Sinauer Associates, Sunderland, MA, 2000.

Margoliash, E. Primary structure and evolution of cytochrome *c. Proc. Natl. Acad. Sci. USA* **50,** 672–679 (1963).

Mayr, E. *The Growth of Biological Thought. Diversity, Evolution, and Inheritance.* Belknap Harvard, Cambridge, MA, 1982.

Moritz, C., and Hillis, D. M. Molecular systematics: Context and controversies. In D. M. Hillis, C. Moritz, and B. K. Mable (Eds.) *Molecular Systematics.* Sinauer Associates, Sunderland, MA, 1996, pp. 1–13.

Nei, M. *Molecular Evolutionary Genetics.* Columbia University Press, New York, 1987, pp. 39–63.

Nei, M. Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.* **30,** 371–403 (1996).

Nuttall, G. H. F. Blood Immunity and Blood Relationship. Cambridge University Press, Cambridge, 1904.

Ohno, S. *Evolution by Gene Duplication.* Springer-Verlag, New York, 1970.

Philippe, H., and Laurent, J. How good are deep phylogenetic trees? *Curr. Opin. Genet. Dev.* **8,** 616–623 (1998).

Posada, D., and Crandall, K. A. MODELTEST: Testing the model of DNA substitution. *Bioinformatics* **14,** 817–818 (1998).

Saitou, N., and Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4,** 406–425 (1987).

Schmidt, H. A., Strimmer, K., Vingron, M., and von Haeseler, A. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18,** 502–504 (2002).

Simpson, G. G. *The Meaning of Evolution: A Study of the History of Life and of Its Significance for Man.* Yale University Press, New Haven, 1952.

Strimmer, K., and von Haeseler, A. Quartet puzzling: A quartet maximun likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13,** 964–969 (1996).

Swofford, D. L., Olsen, G. J., Waddell, P. J., and Hillis, D. M. Molecular systematics: Context and controversies. In D. M. Hillis, C. Moritz, and B. K. Mable (Eds.), *Molecular Systematics,* 2nd ed. Sinauer Associates, Sunderland, MA, 1996.

Thornton, J. W., and DeSalle, R. Gene family evolution and homology: Genomics meets phylogenetics. *Annu. Rev. Genomics Hum. Genet.* **1,** 41–73 (2000).

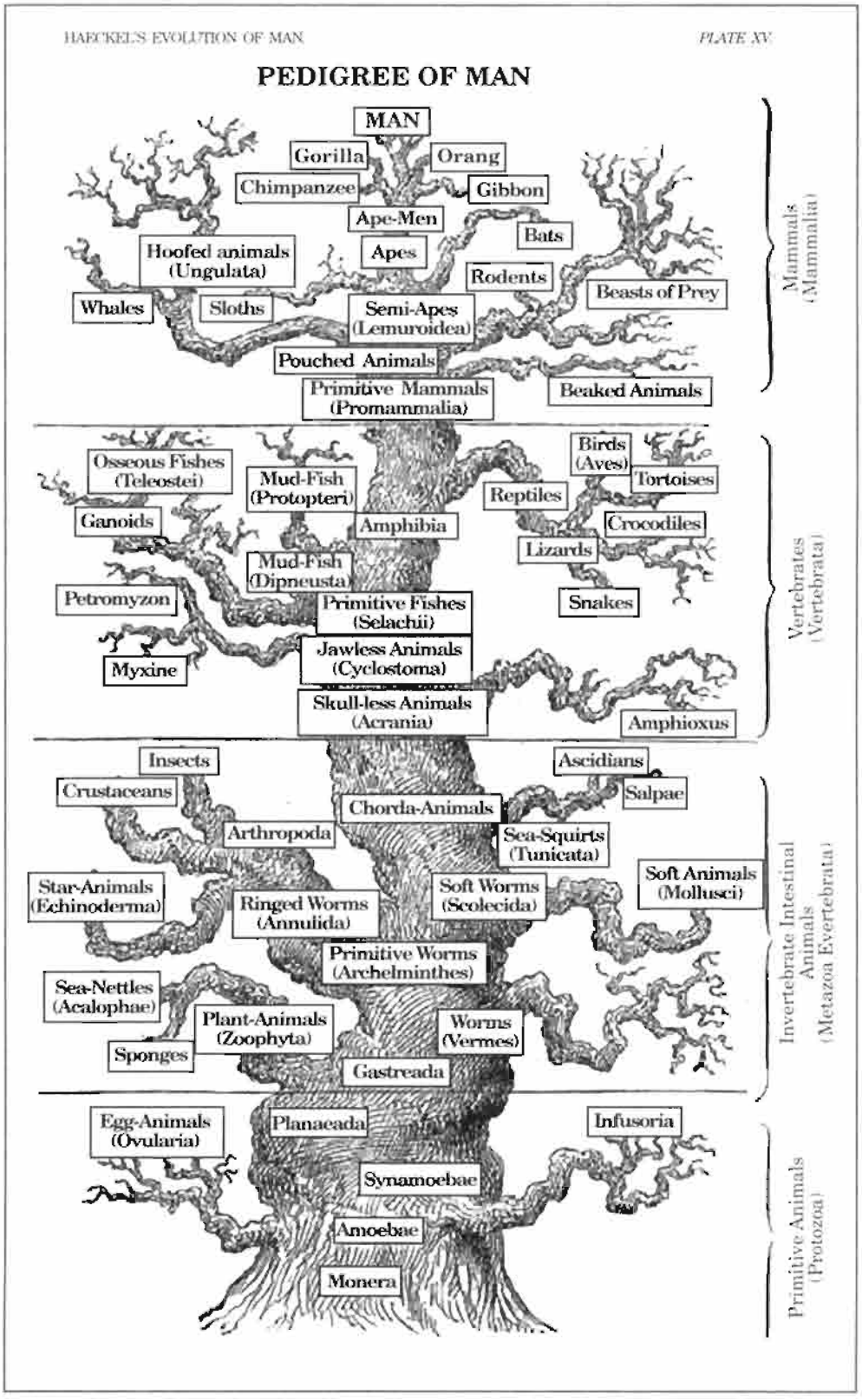
Wilson, E. O. *The Diversity Of Life.* W. W. Norton, New York, 1992.

Zuckerkandl, E., and Pauling, L. Molecular disease, evolution, and genic heterogeneity. In M. Kasha and B. Pullman (Eds.), *Horizons In Biochemistry,* Albert Szent-Gyorgyi Dedicatory Volume. Academic, New York, 1962.

This Page Intentionally Left Blank

Part III

Genome Analysis



The tree of life from Ernst Haeckel (1879). The figure shows mammals (with humans at the top shown ascending from apes), vertebrates, invertebrates, and primitive animals at the bottom, including Monera (bacteria). (Reproduced with permission of the Institute of the History of Medicine, The Johns Hopkins University.)

12

Completed Genomes and the Tree of Life

The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth. The green and budding twigs may represent existing species; and those produced during each former year may represent the long succession of extinct species. . . . The limbs divided into great branches, and these into lesser and lesser branches, were themselves once, when the tree was small, budding twigs; and this connexion of the former and present buds by ramifying branches may well represent the classification of all extinct and living species in groups subordinate to groups. . . . From the first growth of the tree, many a limb and branch has decayed and dropped off, and these lost branches of various sizes may represent those whole orders, families, and genera which have now no living representatives, and which are known to us only from having been found in a fossil state. . . . As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all a feebler branch, so by generation I believe it has been with the Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications.

–Charles Darwin, *The Origin of Species* (1859)

INTRODUCTION

A genome is the collection of DNA that comprises an organism. Each individual organism's genome contains the genes and other DNA elements that ultimately

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

define its identity. Genomes range in size from the smallest viruses, which encode fewer than 10 genes, to eukaryotes such as humans that have billions of base pairs of DNA encoding tens of thousands of genes.

The recent sequencing of genomes from all branches of life—including viruses, prokaryotes, fungi, a nematode, a plant, and humans—presents us with an extraordinary moment in the history of biology. By analogy, this situation resembles the completion of the periodic table of the elements in the nineteenth century. As it became clear that the periodic table could be arranged in rows and columns, it became possible to predict the properties of individual elements. A logic emerged to explain the properties of the elements. But it still took another century to grasp the significance of the elements and to realize the potential of the organization inherent in the periodic table.

Today we have sequenced the DNA from hundreds of genomes, and we are now searching for a logic to explain their organization and function. This process will take decades. A variety of tools must be applied, including bioinformatics approaches, biochemistry, genetics, and cell biology.

This chapter introduces the tree of life and the sequencing of genomes. We will then proceed to assess the progress in studying the genomes of viruses (Chapter 13); prokaryotes (bacteria and archaea) (Chapter 14); fungi, including the yeast *Saccharomyces cerevisiae* (Chapter 15); an assortment of eukaryotes from parasites to primates (Chapter 16); and finally the human genome (Chapters 17 and 18). For definitions of several key terms related to the tree of life, see Table 12.1.

TABLE 12-1 Nomenclature for Tree of Life

Name refers to the name adopted in this book. See Woese et al. (1990).

Name	Synonym(s)	Definition
Archaea (singular: archaeon)	Archaebacteria	One of the three "urkingdoms" or "domains" of life
Bacteria	Eubacteria; Monera (obsolete name)	One of the three "urkingdoms" or "domains" of life; unicellular organisms characterized by lack of a nuclear membrane
Eukaryotes	Eucarya	One of the three "urkingdoms" or "domains" of life; cells characterized by a nuclear membrane
Microbe	–	Microorganisms that cause disease in humans; microbes include bacteria and eukaryotes such protozoa and fungi
Microorganism	–	Unicellular life forms of microscopic size, including bacteria, archaea, and some eukaryotes
Progenote	Last universal common ancestor	The ancient, unicellular life form from which the three domains of life are descended
Prokaryotes	Procaryotes; formerly synonymous with bacteria	Organism lacking a nuclear membrane; bacteria and archaea

Brief History of Systematics

Throughout recorded history, philosophers and scientists have grappled with questions regarding the diversity of life on Earth (Mayr, 1982). Aristotle (384–322 B.C.E.) was an active biologist, describing over 500 species in his zoological works. He did not create a general classification scheme for life, but he did describe animals as "blooded" or "bloodless" in his *Historia animalium*. [Eventually, Lamarck (1744–1829) renamed these categories "vertebrates" and "invertebrates."] Aristotle's division of animals into genera and species provides the origin of the taxonomic system we use today.

The greatest advocate of this binomial nomenclature system of genus and species for each organism was the Swedish naturalist Carl Linnaeus (1707–1778). Linnaeus also introduced the notion of the three kingdoms, Animaliae, Plantae, and Mineraliae; in his hierarchical system the four levels were class, order, genus, and species. Ernst Haeckel (1834–1919), who described over 4000 new species, enlarged this system. He described life as a continuum from mere complex molecules to plants and animals, and he described the Moner as formless clumps of life. The monera were later named bacteria. By the end of the 1960s the work of Haeckel, Copeland, Whittaker, and many others led to the standard five-kingdom system of life: animals, plants, single-celled protists, fungi, and monera. An example of the tree of life, from an 1879 book by Haeckel, is shown in the frontis to this chapter. Meanwhile in parallel to the five-kingdom system, Edouard Chatton (1937) made the distinction between prokaryotes (bacteria that lack nuclei) and eukaryotes (organisms with cells that have nuclei).

A species is a group of similar organisms that only breed with one another, under normal conditions. A genus may consist of between one and hundreds of species.

The tree of life was rewritten in the 1970s and 1980s by Carl Woese and colleagues (Fox et al., 1980; Woese, 1998; Woese et al., 1990). They studied a group of prokaryotes that were presumed to be bacteria because they were single-celled life forms that lack a nucleus. The researchers sequenced small-subunit ribosomal RNAs (SSU rRNA) and performed phylogenetic analyses. This revealed that archaea are as closely related to eukaryotes as they are to bacteria. A phylogenetic analysis of SSU rRNA sequences, which are present in all known life forms, provides one version of the tree of life (Fig. 12.1). There are three main branches. While the exact root of the tree is not known, the deepest branching bacteria and archaea are thermophiles, suggesting that life may have originated in a hot environment.

Viruses do not meet the definition of living organisms, and thus they are excluded from this tree. Although they replicate and evolve, viruses only survive by commandeering the cell of a living organism (see Chapter 13).

History of Life on Earth

Our recent view of the tree of life (Fig. 12.1) is accompanied by new interpretations of the history of life on Earth. All life forms share a common origin and are part of the tree of life. A species has an average half-life of 1–10 million years (Graur and Li, 2000), and more than 99% of all species that ever lived are now extinct (Wilson, 1992). In principle, there is one single tree of life that accurately describes the evolution of species. The object of phylogeny is to try to deduce the correct trees both for species and for homologous families of genes and proteins. Another object of

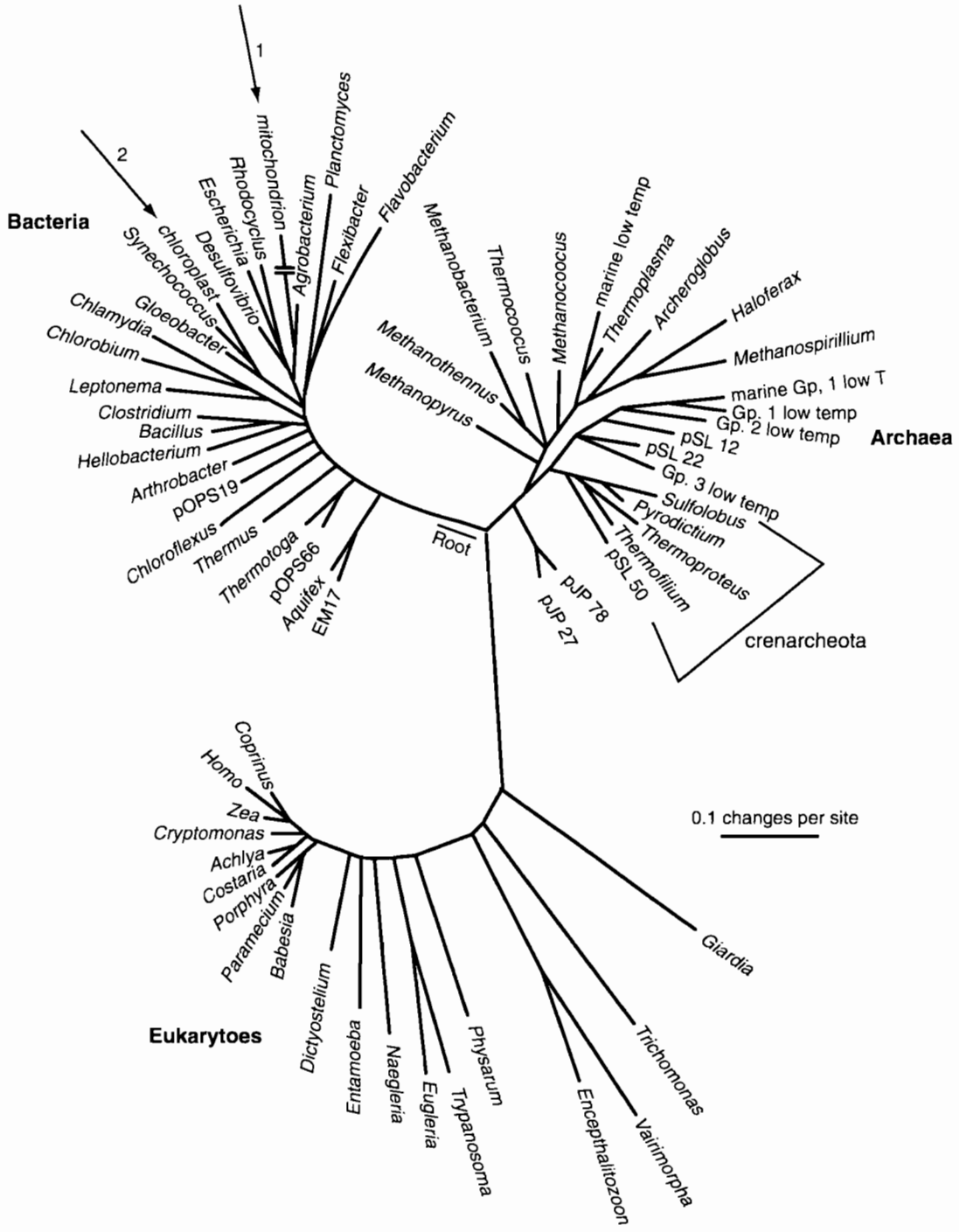


FIGURE 12.1. *A global tree of life, based upon phylogenetic analysis of small-subunit rRNA sequences (modified from Pace, 2001). Life is thought to have originated at least 3.8 BYA in an anaerobic environment. The primordial life form (progenote) displayed the defining features of life (self-replication and evolution). Used with permission.*

phylogeny is to infer the time of divergence between organisms since the time they last shared a common ancestor.

An overview of the history of life is shown in Figure 12.2. The earliest evidence of life is from about 4 billion years ago (BYA), just 0.5 billion years after the formation of Earth. This earliest life was centered on RNA (rather than DNA or protein;

reviewed in Joyce, 2002). Earth's atmosphere was anaerobic throughout much of early evolution, and this early life form was presumably a unicellular prokaryote. The first fossil evidence of life is dated about 3.8 BYA. The last common ancestor of life, predating the divergence of the lineage that leads to modern bacteria and modern archaea, was probably a hyperthermophile. This is suggested by the deepest branching organisms of trees (see Fig. 12.1), such as the bacterium *Aquifex* and the hyperthermophilic crenarcheota (Chapter 14). Eukaryotes appeared between 2 and 3 BYA and remained unicellular until almost 1 BYA. Approximately 1.5 BYA, plants and animals diverged, as did fungi from the lineage that gave rise to metazoans (animals) (see Fig. 16.22). The most recent billion years of life has seen the evolution of an enormous variety of multicellular organisms. The so-called Cambrian explosion of 550 million years ago (MYA) witnessed a tremendous increase in the diversity of animal life forms. In the past 250 million years, the continents coalesced into a giant continent, Pangaea (Fig. 12.3). When Pangaea separated into northern and southern supercontinents (Laurasia and Gondwana), this created natural barriers to reproduction and influenced subsequent evolution of life. By 60 MYA, the dinosaurs were extinct, and the mammalian radiation was well underway.

Multicellular organisms evolved independently many times. A variety of multicellular bacteria evolved several billion years ago, allowing selective benefits in feeding and in dispersion from predators (Kaiser, 2001).

The lines leading to modern *Homo sapiens,* chimpanzees, and bonobos diverged about 5 MYA. The earliest human ancestors include "Lucy," the early *Australepithecus,* and early hominids used stone tools over 2 MYA. Further features of recent historical interest are indicated in Figure 12.2.

Molecular Sequences as Basis of Tree of Life

In past decades and centuries, the basis for proposing models of the tree of life was primarily morphology. Thus Linnaeus divided animals into six classes (mammals, birds, fish, insects, reptiles, and worms), subdividing mammals according to features of their teeth, fish according to their fins, and insects by their wings. Early microscopic studies revealed that bacteria lack nuclei, allowing a fundamental separation of bacteria from the four other kingdoms of life. Bacteria could be classified based upon biochemical properties [e.g., by Albert Jan Kluyver (1888–1956)], and from a morphological perspective bacteria can be classified into several major groups. However, such criteria are insufficient to appreciate the dazzling diversity of millions of microbial species. Thus physical criteria were unavailable by which to discover archaea as a distinct branch of life.

The advent of molecular sequence data has transformed our approach to the study of life. Such data were generated beginning in the 1950s and 1960s, and by 1978 Dayhoff's Atlas used several hundred protein sequences as the basis for PAM matrices (Chapter 3). With the rapid rise in available DNA sequences of the past several years, phylogenetic analyses are now possible based upon both phenotypic characters and based upon gene sequences. The most widely used sequences are SSU rRNA molecules, which are present across virtually all extant life forms. The slow rate of evolution of SSU rRNAs and their convenient size makes them appropriate for phylogenetic analyses. Genome-sequencing efforts are now reshaping the field of evolutionary studies, providing thousands of DNA and protein sequences for phylogenetic trees.

The European Small Subunit Ribosomal RNA Database (Van de Peer et al., 2000; Wuyts et al., 2002) (▸http://rrna.uia.ac.be/ssu/ or ▸http://oberon.rug.ac.be:8080/rRNA/) currently contains over 20,000 SSU rRNA sequences (last updated September 2001). About 600 are archaeal species, 12,000 bacteria, and 6500 eukaryotes. You can use sequences from this database to generate phylogenetic trees or view trees at the website.

Over 100 prokaryotic genomes have now been sequenced (see below and Chapter 14). We are now beginning to appreciate lateral gene transfer (Chapter 14), a phenomenon in which a species does not acquire a particular gene by descent from

FIGURE 12.2. *History of life on the planet. Sources include Kumar and Hedges (1998), Hedges et al. (2001), and Benton and Ayala (2003).*

an ancestor. Instead, it acquires the gene horizontally (or laterally) from another unrelated species. Thus genes can be exchanged between species (Eisen, 2000). As a consequence, the use of different individual genes in molecular phylogeny often results in distinctly different tree topologies. Because of the phenomena of lateral gene transfer and gene loss, it might never be possible to construct a single tree of life that reflects the evolution of life on the planet (Wolf et al., 2002).

Role of Bioinformatics in Taxonomy

The field of bioinformatics is concerned with the use of computer algorithms and computer databases to elucidate the principles of biology. The domain of

FIGURE 12.2. *(Continued)*

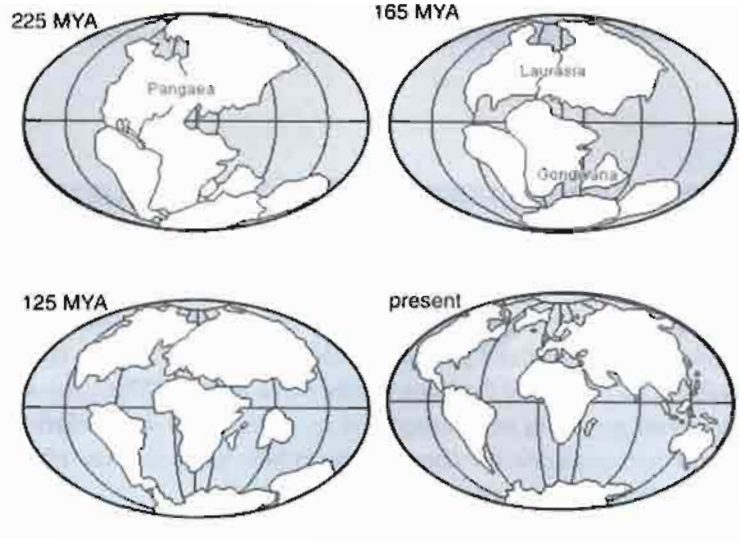


FIGURE 12.3. *Geological history of the earth from 225 MYA. At that time, there was one supercontinent, Pangaea. By 165 MYA, Pangaea had separated into Laurasia (modern Asia and North America) and Gondwana (modern Africa and South America). At 125 MYA Laurasia and Gondwana had both begun separations that led to the present divisions among continents.*

bioinformatics includes the study of genes, proteins, and cells in the context of organisms across the tree of life. Some have advocated a web-based taxonomy intended to catalog an inventory of life (Blackmore, 2002; Pennisi, 2001). Several projects attempt to create a tree of life (see sidebar). Others suggest that while web-based initiatives are useful, the current system is adequate: Zoological, botanical, or other specimens are collected, named, and studied according to guidelines established by international conventions (Knapp et al., 2002).

The Convention on Biological Diversity (▸ http://www.biodiv.org/) and the Global Biodiversity Information Facility (▸ http://www.gbif.net/) are examples of organizations that address issues of global biodiversity.

The Tree of Life is at ▸ http://www.panspermia.org/tree.htm and the Tree of Life Web Project (created by David R. Maddison) is at ▸ http://tolweb.org/tree/phylogeny.html.

Genome-Sequencing Projects: Historical Overview

The advent of DNA-sequencing technologies in the 1970s, including Fred Sanger's dideoxynucleotide methodology, enabled large-scale sequencing projects to be performed. This chapter provides a brief history of genome-sequencing projects, including the completion of the genomic sequence of the first free-living organism in 1995, *Haemophilus influenzae*. By 2001, a draft sequence of the human genome was reported by two groups. The most remarkable feature of current efforts to determine the sequence of complete genomes is the dramatic increase in data that are collected each year (Fig. 2.1). The ability to sequence first millions and now billions of nucleotides of genomic DNA presents the scientific community with unprecedented opportunities and challenges.

Several themes have emerged in the past several years (Webb, 2001):

- The amount of sequence data that are generated continues to accelerate rapidly.
- For many genomes, even unfinished genomic sequence data—that is, versions of genomic sequence that include considerable gaps and sequencing errors—are immediately available and useful to the scientific community. We will also see that a finished sequence (defined below) provides substantially better descriptions of genome features than does an unfinished sequence.
- The value of comparative genome analysis is now appreciated for solving problems such as identifying protein-coding genes in human and mouse or differences in virulent and nonvirulent strains of pathogens (Koonin, 2001). Comparative analyses will also be useful to define gene regulatory regions and the evolutionary history of species through the analysis of conserved DNA elements.

Brief Chronology of Genome-Sequencing Projects

Overview: 1977–Present

The progress in completing dozens of genome-sequencing projects has been rapid, and we can expect the pace to accelerate in the future. A chronological overview is helpful to provide a framework for these events. When the sequencing of the first bacterial genomes was completed in 1995, there were relatively few other genome sequences available for comparison. There are now over 1000 completed genomes available (including organellar genomes).

TABLE 12-2 Web-Based Resources for Completed Genome Projects

Most of these sites contain links to additional, related resources.

Resource	Description	URL
EBI	European Bioinformatics Institute	▸http://www.ebi.ac.uk/genomes/
GNN	Genome News Network; includes photographs of organisms	▸http://gnn.tigr.org/sequenced.genomes/genome.guide.p1.shtml
GOLD	Genomes Online Database	▸http://ergo.integratedgenomics.com/GOLD/
Infobiogen	Complete microbial genomes	▸http://www.infobiogen.fr/doc/data/complete_genome.html
NCBI	Entrez at National Center for Biotechnology Information	▸http://www.ncbi.nlm.nih.gov/Entrez/
PEDANT	Protein Extraction, Description and Analysis Tool	▸http://pedant.gsf.de/
TIGR	The Institute for Genomic Research	▸http://www.tigr.org/tdb

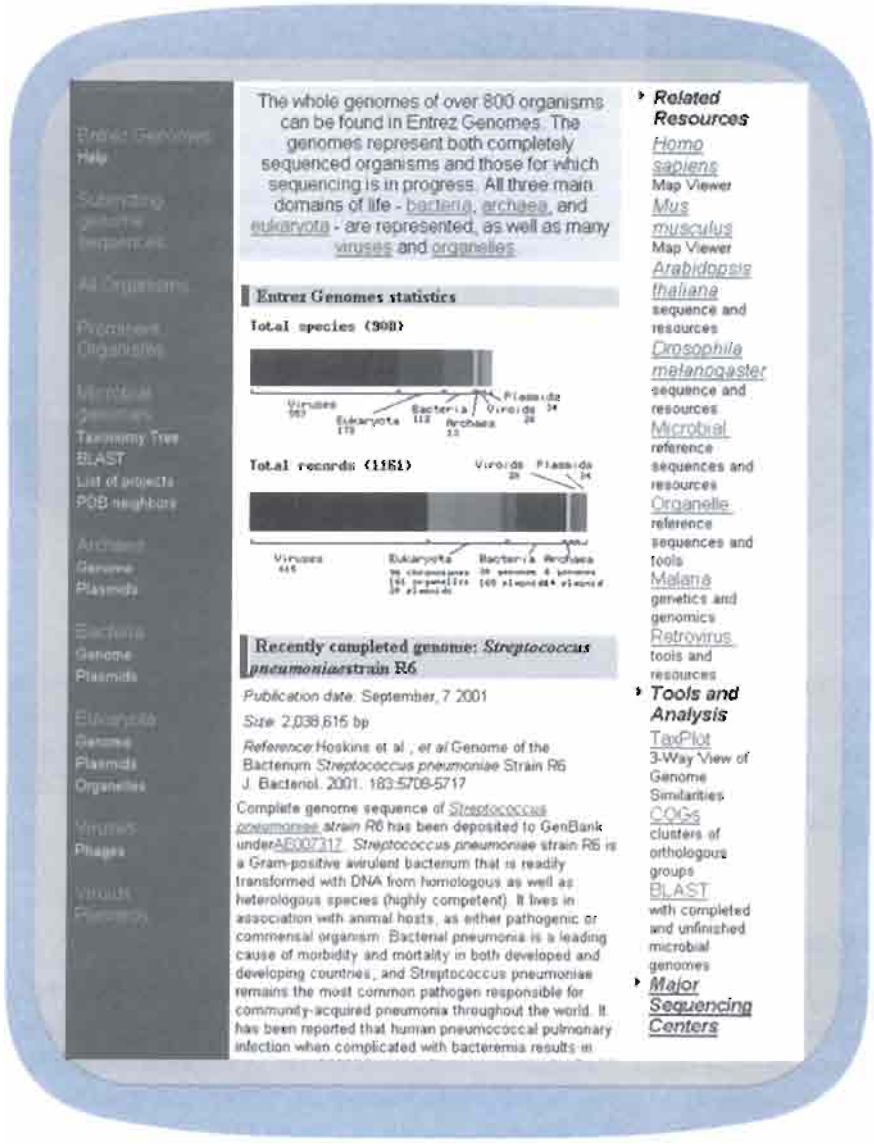


FIGURE 12.4. *The Entrez Genome site of NCBI includes links to molecular sequence data from over 100,000 species (left sidebar). Related resources for selected organisms are also provided as well as several tools for genome analysis (right sidebar). The nucleotide sequence is available for over 900 genomes. This page is accessible from* ▸http://www.ncbi.nlm.nih.gov/Entrez/.

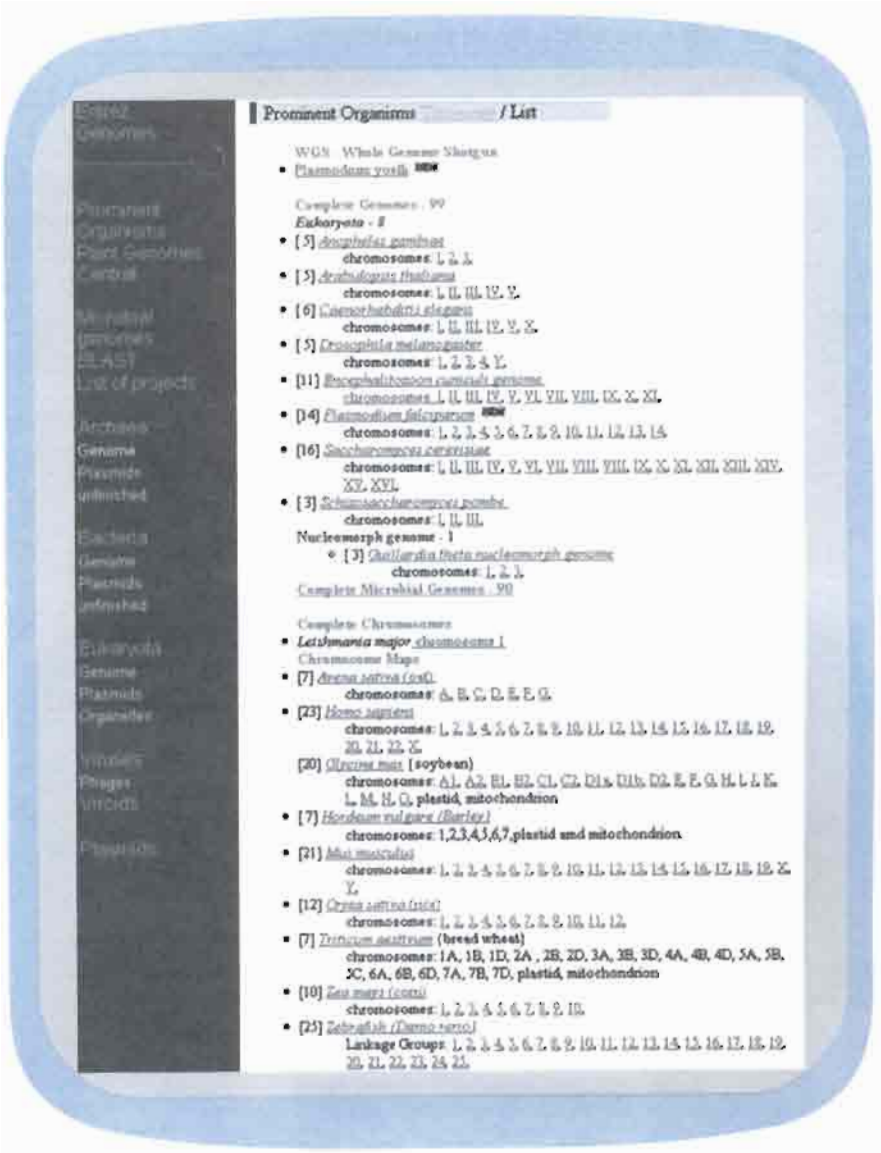


FIGURE 12.5. The Entrez Genome prominent organism page provides links to sequenced genomes, including microbial genomes and chromosome maps of selected eukaryotes.

We will introduce three main web resources for the study of genomes in this and the following chapters: EBI, TIGR, and NCBI (these and other sites are listed in Table 12.2). We begin with the genomes section of Entrez at NCBI (Fig. 12.4). Clicking "Prominent Organisms" on the left sidebar gives an overview of the sites on eukaryotic, bacterial, and archaeal genomes (Fig. 12.5).

First Viral Genome (1977)

Fred Sanger and colleagues sequenced the genome of bacteriophage ϕX174 (Sanger et al., 1977). They developed several DNA-sequencing techniques, including the dideoxynucleotide chain termination procedure. Bacteriophage ϕX174 is 5386 base pairs (bp) encoding 11 genes (see GenBank accession J02482). A graphical depiction of a portion of this viral genome is shown in Figure 12.6. At the time, the most

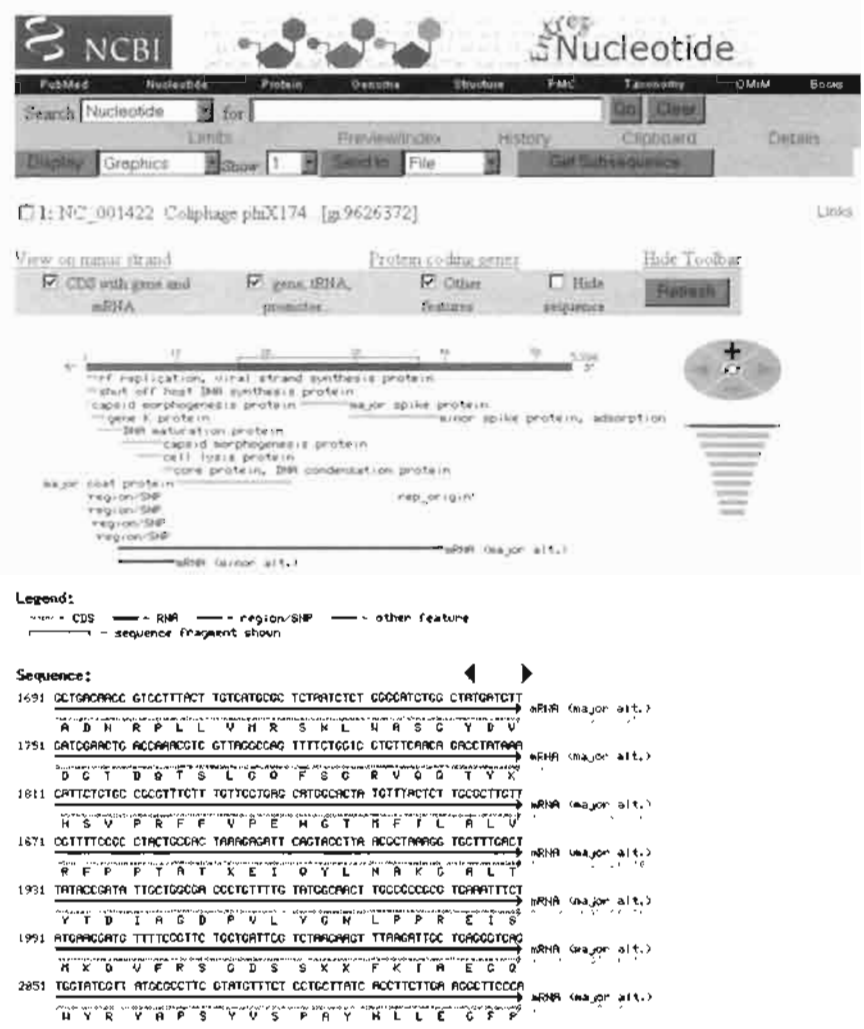


FIGURE 12.6. *Portion of the Entrez nucleotide record for bacteriophage ϕX174. This format was obtained by viewing the entry for accession J02482 in the graphics display format. This provides an overview of the predicted open reading frames (ORFs). A portion of the nucleotide sequence and corresponding predicted proteins is shown at the bottom.*

surprising result was the unexpected presence of overlapping genes that are transcribed on different reading frames.

First Eukaryotic Organellar Genome (1981)

The first complete organellar genome to be sequenced was the human mitochondrion (Anderson et al., 1981). The genome is characterized by extremely little noncoding DNA. The great majority of metazoan (i.e., multicellular animal) mitochondrial genomes are about 15–20 kb (kilobase) circular genomes. The human mitochrondrial genome is 16,568 bp (base pairs) and encodes 13 proteins, 2 ribosomal RNAs, and 22 transfer RNAs. It can be accessed through the NCBI Entrez genome site (Fig. 12.7). This circular diagram is clickable, allowing you to access the individual genes (Fig. 12.8). Thus, the DNA and corresponding protein sequences of all the mitochondrial genes are easily accessible.

Today, there are over 400 completed mitochondrial genome sequences. Several of these are listed in Table 12.3. This table also lists several exceptionally large cases. While the largest sequenced mitochrondrial genome is that of the thale cress *Arabidopsis thaliana* (367 kb), several plants reportedly have even larger mitochondrial genomes. Thus, there is a tremendous diversity of mitochondrial genomes (Lang

In dideoxynucleotide sequencing, a template DNA is copied with a polymerase in the presence of a radioactive nucleotide (e.g., [^{32}P]dATP, used as a tracer to visualize the newly synthesized DNA strand), four nucleotides (dATP, dGTP, dCTP, dTTP), and a small amount of dideoxynucleotide [either ddATP, ddGTP, ddCTP, or ddTTP (dideoxythymidine triphosphate), used in four separate reactions]. The dideoxynucleotides cause chain termination, leading to products with the same 5′ end but differing in length at the 3′ end. These products can be electrophoresed on a polyacrylamide gel, and the base composition can then be inferred.

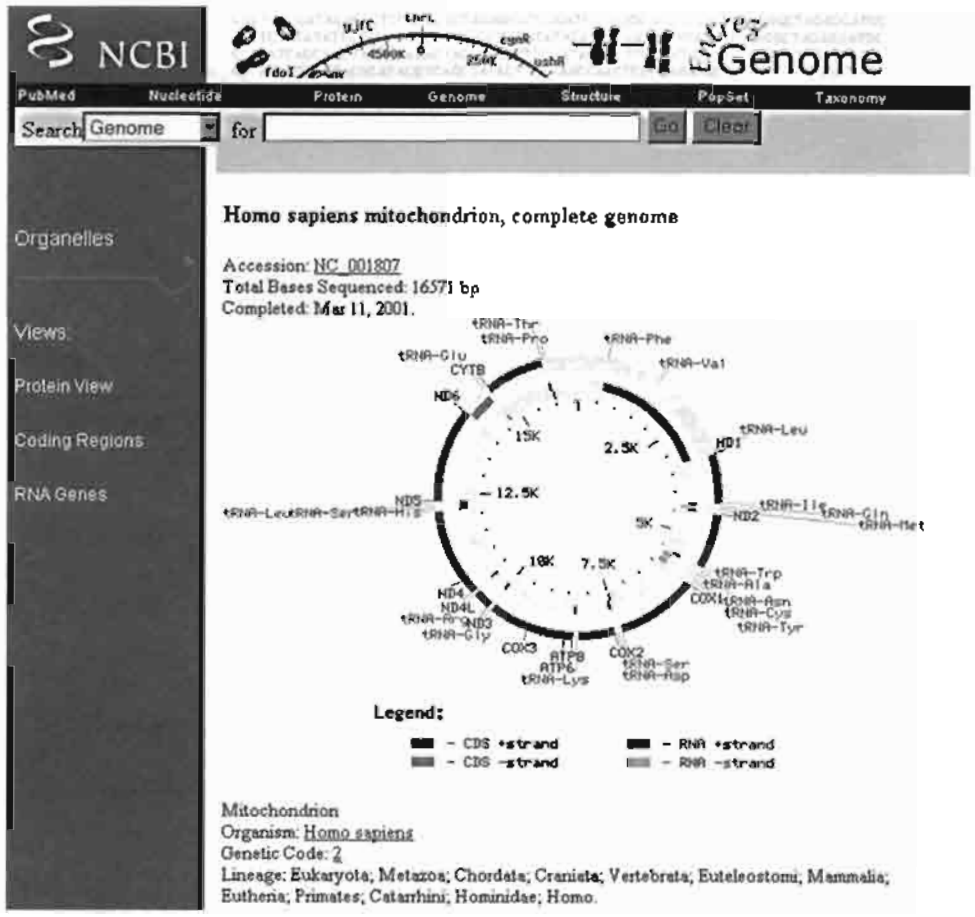


FIGURE 12.7. *NCBI entry for the human mitochondrial genome (NC.001807; 16,568 bp). Further details of the sequence can be obtained by clicking the main image (see Fig. 12.8 a) or by clicking the "Coding Regions" link on the left sidebar (see Fig. 12.8 b).*

Information on mitochondrial genomes is available at ▸http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/mitostat.html.

et al., 1999). Molecular phylogenetic approaches suggest that mitochondria are descendants of an endosymbiotic α-proteobacterium, although it is possible that the origin of mitochondria in eukaryotes was coincident with the evolution of the nuclear genome (Lang et al., 1999).

First Chloroplast Genomes (1986)

We will discuss chloroplasts and other plastids in the plant section of Chapter 16.

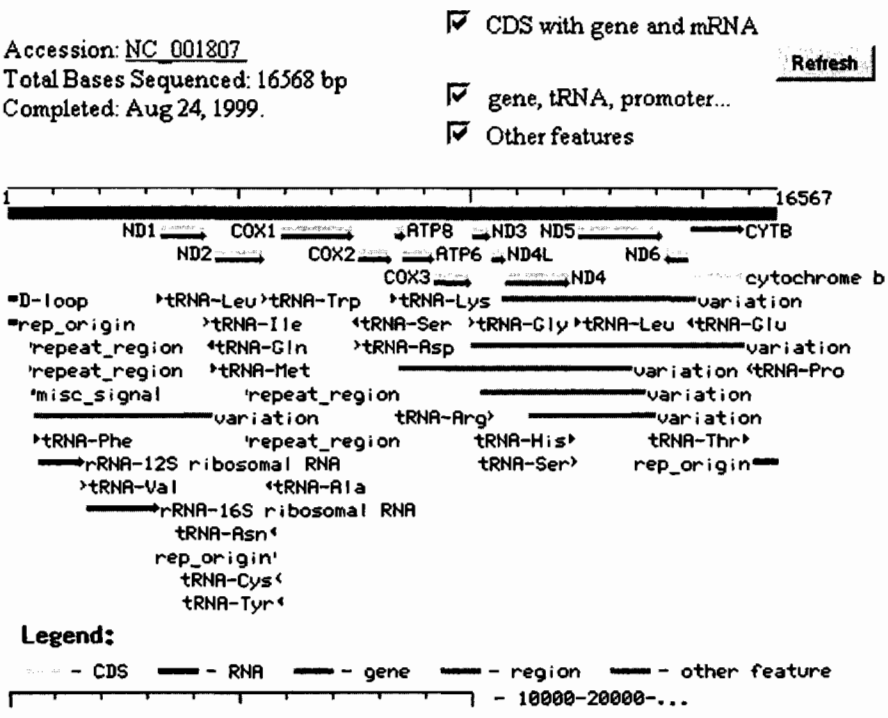
The first chloroplast genomes were reported (*Nicotiana tabacum*; Shinozaki et al., 1986), followed by the liverwort *Marchantia polymorpha* (Ohyama et al., 1986). Most plant chloroplast genomes are 120,000–200,000 bp in size. There are other chloroplast-like organelles in eukaryotic organisms. Unicellular protozoan parasites of the phylum Apicomplexa, such as *Toxoplasma gondii* (Table 12.4), have smaller plastid genomes.

First Eukaryotic Chromosome (1992)

The first eukaryotic chromosome was sequenced in 1992: chromosome III of the budding yeast *S. cerevisiae* (Oliver et al., 1992). There were 182 predicted open reading frames (for proteins larger than 100 amino acids), and the size of the sequenced

(*a*)

Homo sapiens mitochondrion, complete genome [1 .. 16567]



(*b*)

Homo sapiens mitochondrion, complete genome

Save the report below in Table format.

◆ - GenBank record including protein ◆ - DNA region in flatfile format ● - DNA and protein in FASTA format

	Location	Strand	Length	PID	Gene	Synonym	Product
◆ ◆ ◆	3306..4262	+	319	5835388	ND1		NADH dehydrogenase subunit 1
◆ ◆ ◆	4469..5512	+	348	5835389	ND2		NADH dehydrogenase subunit 2
◆ ◆ ◆	5903..7444	+	514	5835390	COX1		cytochrome c oxidase subunit I
◆ ◆ ◆	7585..8268	+	228	5835391	COX2		cytochrome c oxidase subunit II
◆ ◆ ◆	8365..8571	+	69	5835392	ATP8		ATP synthase F0 subunit 8
◆ ◆ ◆	8526..9206	+	227	5835393	ATP6		ATP synthase F0 subunit 6
◆ ◆ ◆	9206..9988	+	261	5835394	COX3		cytochrome c oxidase subunit III
◆ ◆ ◆	10058..10405	+	116	5835395	ND3		NADH dehydrogenase subunit 3
◆ ◆ ◆	10469..10765	+	99	5835396	ND4L		NADH dehydrogenase subunit 4L
◆ ◆ ◆	10759..12138	+	460	5835397	ND4		NADH dehydrogenase subunit 4
◆ ◆ ◆	12336..14147	+	604	5835398	ND5		NADH dehydrogenase subunit 5
◆ ◆ ◆	14148..14672	-	175	5835399	ND6		NADH dehydrogenase subunit 6
◆ ◆ ◆	14746..15880	+	378	6137797	CYTB		cytochrome b

FIGURE 12.8. (a) *Detailed information on the human mitochondrial genome (NC_001807) is available by clicking on a specific gene (e.g.,* COX1; *Fig. 12.7).* (b) *A link from the left sidebar of the mitochondrial genome page provides a list of all mitochrondrial protein-coding genes.*

DNA was 315 kb. Of the 182 open reading frames that were identified, only 37 corresponded to previously known genes, and 29 showed similarity to known genes. We will explore this genome in Chapter 15.

Complete Genome of Free-Living Organism (1995)

The first genome of a free-living organism to be completed was the bacterium *H. influenzae* Rd (Fleischmann et al., 1995). Its size is 1,830,138 bp [i.e., 1.8 Mb (megabase pairs)]. This organism was sequenced at The Institute for Genomic Research (TIGR; ▸http://www.tigr.org) using the whole-genome shotgun sequencing and assembly strategy (see below).

We describe this bacterial genome as derived from a "free-living organism" to distinguish it from a viral genome or an organellar genome. Viruses (Chapter 13) exist on the borderline of the definition of life, and organellar genomes are usually derived from bacteria that are no longer capable of independent life.

TABLE 12-3 Selected Mitochondrial Genomes Arranged by Size

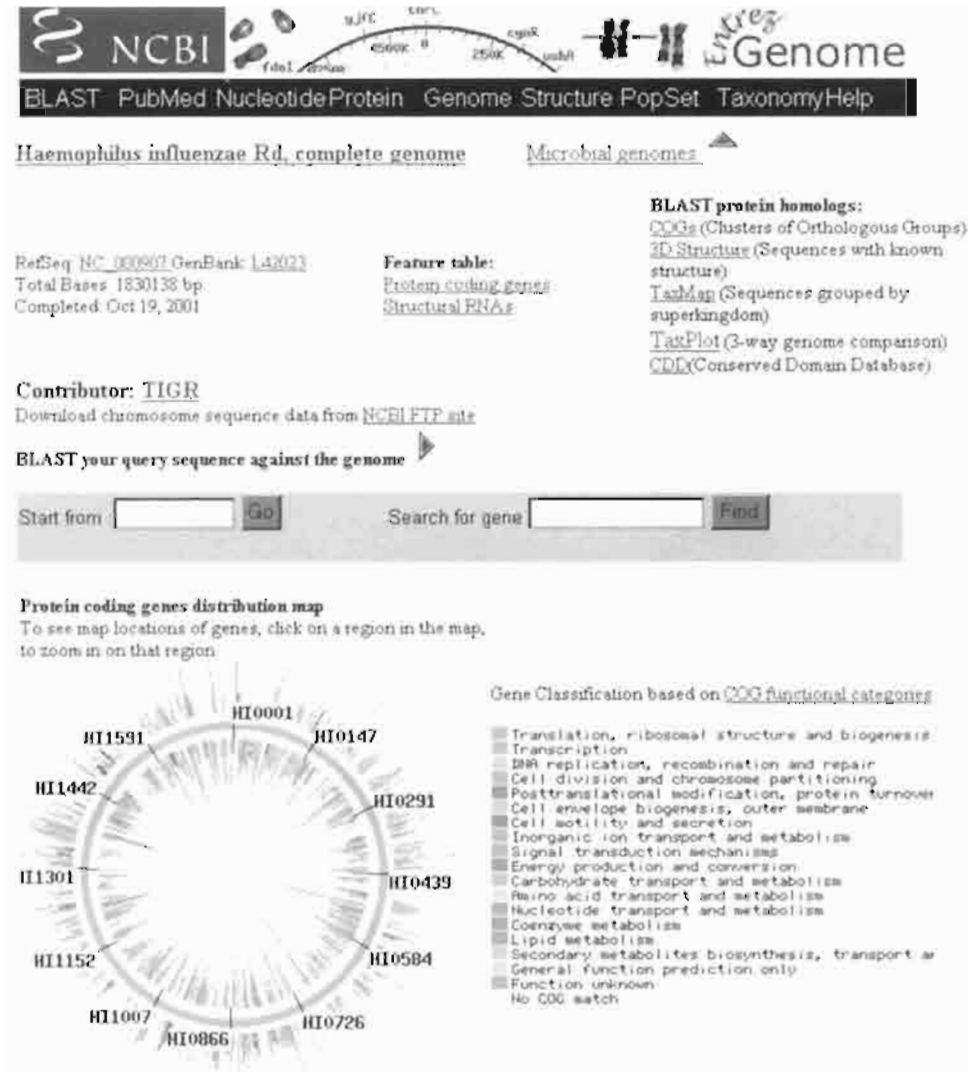
As of June 2003, 357 metazoan (multicellular animal) organellar genomes have been sequenced, 23 fungi, 10 plants, and 23 other eukaryotes (▸http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/mitostat.html)

Kingdom	Species	Accession	Size (bp)
Eukaryote	*Plasmodium falciparum* (malaria parasite)	NC_002375	5,967
Metazoa (Bilateria)	*Caenorhabditis elegans* (worm)	NC_001328	13,794
Plant (Chlorophyta)	*Chlamydomonas reinhardtii* (green alga)	NC_001638	15,758
Metazoa (Bilateria)	*Mus musculus*	NC_001569	16,295
	Pan troglodytes (chimpanzee)	NC_001643	16,554
Metazoa	*Homo sapiens*	NC_001807	16,568
Metazoa (Cnidaria)	*Metridium senile* (sea anenome)	NC_000933	17,443
Metazoa (Bilateria)	*Drosophila melanogaster*	NC_001709	19,517
Fungi (Ascomycota)	*Schizosaccharomyces pombe*	NC_001326	19,431
	Candida albicans	NC_002653	40,420
Eukaryote (stramenopiles)	*Pylaiella littoralis* (brown alga)	NC_003055	58,507
Fungi (Chytridiomycota)	*Rhizophydium* sp. 136	NC_003053	68,834
Eukaryote	*Reclinomonas americana* (protist)	NC_001823	69,034
Fungi (Ascomycota)	*Saccharomyces cerevisiae*	NC_001224	85,779
Plant (Streptophyta)	*Arabidopsis thaliana*	NC_001284	366,923
Plant	*Zea mays* (corn)	Fauron et al. (1995)	700,000
	Cucumis melo	Lilly and Havey (2001)	2,400,000

To study this genome in NCBI, go to Entrez, click "genome," and then click "bacteria genome" from the left sidebar. Note that there are over 100 complete bacterial genomes listed; scroll down to *H. influenzae* and click on the accession number. [Note that you can also see the RefSeq accession number, NC_000907; the genome size (1,830,138 bp); and the date entered, July 25 1995.] The page for this organism contains a wealth of information on the genes and encoded proteins as well as the predicted functional classification of the proteins (Fig. 12.9). This classification scheme, Clusters of Orthologous Genes (COG), will be discussed in Chapter 14. The lineage (Fig. 12.9, bottom) shows you that this is a bacterium in the gamma division. As with the mitochondrial genome (Fig. 12.7), the circular diagram of the *H. influenzae* genome is clickable to allow a detailed study of its genes and proteins (Fig. 12.10).

TABLE 12-4 Selected Chloroplast Genomes

Species	Common Name	Accession	Size (bp)
Arabidopsis thaliana	Thale cress	NC_000932	154,478
Guillardia theta	Red alga	NC_000926	121,524
Marchantia polymorpha	Liverwort; moss	NC_001319	121,024
Nicotiana tabacum	Tobacco	NC_001879	155,939
Oryza sativa	Rice	NC_001320	134,525
Porphyra purpurea	Red alga	NC_000925	191,028
Toxoplasma gondii	Apicomplexan parasite	NC_001799	34,996
Zea mays	corn	NC_001666	140,384



Organism: Haemophilus influenzae Rd
Genetic Code: 11
Lineage: Bacteria, Proteobacteria; Gammaproteobacteria; Pasteurellaceae; Haemophilus.

FIGURE 12.9. *Entrez Genome record for* H. influenzae *Rd, the first free-living organism for which the complete genomic sequence was determined. This record is obtained from the Entrez Genome resource by clicking "bacteria" on the left sidebar. The top of this entry includes information such as the accession number and the size of the genome. The entire nucleotide sequence is downloadable here. At top right are several resources for studying the 1709 proteins encoded by this genome. An* H. influenzae*-specific BLAST search is available here. The main part of the entry consists of a color-coded circular representation of the genome, along with a functional classification based on COG (see below). The circular map is clickable (see Fig. 12.10), showing detailed information on the genes and proteins in this genome. At the bottom, the genetic code used by this bacterium is provided as well as the taxonomic lineage. The record also contains literature references (not shown), including the initial report of this genomic sequence by Fleischmann et al. (1995) at The Institute for Genomic Research.*

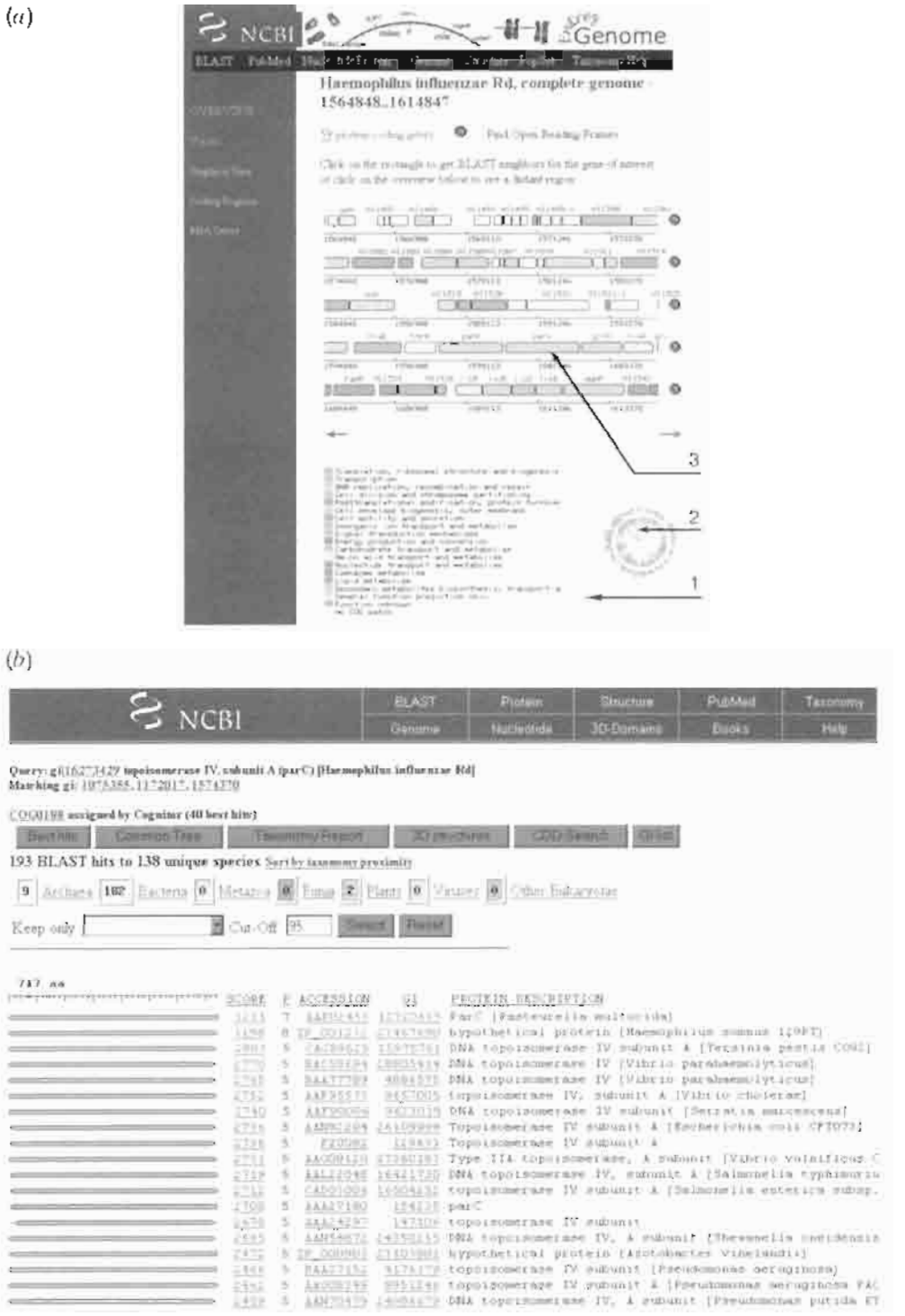


FIGURE 12.10. *Tools to explore completed genomes.* (a) *By clicking on the circular diagram of the* H. influenzae *genome (Fig. 12.9), one can view the protein-coding genes in a detailed view. Proteins are color coded according to the COG scheme (arrow 1) (Table 8.9 on p. 249 and Chapter 14). The region of the genome is indicated (arrow 2). Each protein is clickable.* (b) *By clicking on parC (arrow 3), more information is obtained. A list of alignments to possibly orthologous proteins is shown, including several bacterial, viral, and plant proteins.*

TABLE 12-5 Genome-Sequencing Projects Completed in 1995

Organism	Size (bp)	Accession	Reference
Haemophilus influenzae Rd	1,830,138	NC_000907	Fleischmann et al., 1995
Mycoplasma genitalium	580,074	NC_000908	Fraser et al., 1995

By the end of 1995 (Table 12.5), the complete DNA sequence of a second bacterial genome had been obtained, *Mycoplasma genitalium*. Notably, this is one of the smallest known genomes of any free-living organism. We will explore this and other bacteria in more detail in Chapter 14.

First Eukaryotic Genome (1996)

The complete genome of the first eukaryote, *S. cerevisiae* (a yeast; Chapter 15) (Goffeau et al., 1996), was sequenced by 1996 (Table 12.6). This was accomplished by a collaboration of over 600 researchers in 100 laboratories spread across Europe, North America, and Japan. To find information about this completed genome, go to Entrez genomes, then use the left sidebar to click either "prominent organisms" or "eukaryota genome."

In 1996, TIGR researchers reported the first complete genome sequence for an archaeon, *Methanococcus jannaschii* (Bult et al., 1996). This offered the first opportunity to compare the three main divisions of life, including the overall metabolic capacity of bacteria, archaea, and eukaryotes.

More Bacteria and Archaea (1997)

In 1997, the complete genomic sequences of two archaea were reported (Table 12.7). Of the five bacterial genomes that were reported, the most well-known is that of *Escherichia coli* (Blattner et al., 1997; Koonin, 1997), which has served as a model organism in bacteriology for decades. Its 4.6-Mb genome encodes over 4200 proteins, of which 38% had no identified function at the time. We will explore this further in Chapter 14.

TABLE 12-6 Genome-Sequencing Projects Completed in 1996

The size does not include extrachromosomal elements (for *M. jannaschii*).

Organism	Size	Accession	Reference
Methanococcus jannaschii (A)	1,664,976 bp	NC_000909	Bult et al., 1996
Mycoplasma pneumoniae (B)	816,394 bp	NC_000912	Himmelreich et al., 1996; see Dandekar et al., 2000
Synechocystis PCC6803 (B)	3,573,470 bp	NC_000911	Kaneko et al., 1996
Saccharomyces cerevisiae (E)	12,068 kb	Various, for each chromosome	Goffeau et al., 1996

Abbreviations: A, archaeon; B, bacterium; E, eukaryote.

TABLE 12-7 Genome-Sequencing Projects Completed in 1997

Organism	Size (bp)	Accession	Reference
Archaeoglobus fulgidus (A)	2,178,400	NC_000917	Klenk et al., 1997
Methanobacterium thermoautotrophicum (A)	1,751,377	NC_000916	Smith et al., 1997
Bacillus subtilis (B)	4,214,814	NC_000964	Kunst et al., 1997
Borrelia burgdorferi (B)	910,724	NC_001318	Fraser et al., 1997; Casjens et al., 2000
Escherichia coli (B)	4,639,221	NC_000913	Blattner et al., 1997
Helicobacter pylori 26695 (B)	1,667,867	NC_000915	Tomb et al., 1997; see Alm et al., 1999

Abbreviations: A, archaeon; B, bacterium.

First Genome of Multicellular Organism (1998)

The nematode *Caenorhabditis elegans* was the first multicellular organism to have its genome sequenced—although technically, the sequencing is still not complete (because of the presence of repetitive DNA elements that have been difficult to resolve). The sequence spans 97 Mb and is predicted to encode over 19,000 genes (The C. elegans Sequencing Consortium, 1998).

Two more archaea bring the total to four sequenced genomes by 1998 (Table 12.8). Six more bacterial genomes were also sequenced. The genome of sequence of *Rickettsia prowazekii,* the α-proteobacterium that causes typhus and was responsible for tens of millions of deaths in the twentieth century, is very closely related to the eukaryotic mitochondrial genome (Andersson et al., 1998).

Human Chromosome (1999)

In 1999, the sequence of the euchromatic portion of human chromosome 22 was published (Table 12.9) (Dunham et al., 1999). This was the first human chromosome to be essentially completely sequenced. We will describe the human genome in Chapter 17.

TABLE 12-8 Genome-Sequencing Projects Completed in 1998

Organism	Size	Accession	Reference
Pyrococcus horikoshii OT3 (A)	1,738,505 bp	NC_000961	Kawarabayasi et al., 1998
Aquifex aeolicus (B)	1,551,335 bp	NC_000918	Deckert et al., 1998
Chlamydia trachomatis (B)	1,042,519 bp	NC_000117	Stephens et al., 1998
Chlamydophila pneumoniae (B)	1,230,230 bp	NC_000922	
Mycobacterium tuberculosis (B)	4,411,529 bp	NC_000962	Cole et al., 1998
Rickettsia prowazekii (B)	1,111,523 bp	NC_000963	Andersson et al., 1998
Treponema pallidum (B)	1,138,011 bp	NC_000919	Fraser et al., 1998
Caenorhabditis elegans (E)	97 Mb	AE000001 AE000002 AE000003 AE000004 AE000005 AE000006	The C. elegans Sequencing Consortium, 1998

Abbreviations: A, archaeon; B, bacterium; E, eukaryote.

TABLE 12-9 Genome-Sequencing Projects Completed in 1999

Organism	Size (bp)	Accession	Reference
Aeropyrum pernix (A)	1,669,695	NC_000854	Kawarabayasi et al., 1999
Thermoplasma volvanium GSS1 (A)	1,584,804	NC_002689	Kawashima et al., 1999
Chlamydia pneumoniae (B)	1,229,858	NC_002179	Kalman et al., 1999
Deinococcus radiodurans (B)	2,648,638 412,348	NC_001263 NC_001264	White et al., 1999 White et al., 1999
Helicobacter pylori J99 (B)	1,643,831	NC_000921	Alm et al., 1999; see Tomb et al., 1997
Thermotoga maritima (B)	1,860,725	NC_000853	Nelson et al., 1999
Homo sapiens chromosome 22 (E)	33,792,315	NT_001454, 10 other contigs	Dunham et al., 1999

Abbreviations: A, archaeon; B, bacterium; E, eukaryote.

Fly, Plant, and Human Chromosome 21 (2000)

In this year, the completed genome sequences of the fruit fly *Drosophila melanogaster* and the plant *A. thaliana* were reported, bringing the number of eukaryotic genomes to four (with a yeast and a worm) (Table 12.10). The *Drosophila* sequence was obtained by scientists at Celera Genomics and the Berkeley Drosophila Genome Project (BDGP) (Adams et al., 2000). There are approximately 13,500 annotated genes. *Arabidopsis* is a thale cress of the mustard family. Its compact genome serves as a model for plant genomics.

We will describe these and other eukaryotic genomes in Chapter 16.

Also in the year 2000, human chromosome 21 was the second human chromosome sequence to be reported (Hattori et al., 2000). This is the smallest of the human autosomes. An extra copy of this chromosome causes Down syndrome, the most common inherited form of mental retardation.

Meanwhile, bacterial genomes continued to be sequenced, and many surprising properties emerged. The genome of *Neisseria meningitidis*, which causes bacterial meningitis, contains hundreds of repetitive elements. Such repeats are more typically associated with eukaryotes. The *Pseudomonas aeruginosa* genome is 6.3 Mb, making it the largest of the sequenced bacterial genomes at that time (Stover et al., 2000).

Among the archaea, the genome of *Thermoplasma acidophilum* was sequenced (Ruepp et al., 2001). This organism thrives at 59°C and pH 2. Remarkably, it has undergone extensive lateral gene transfer with *Sulfolobus solfataricus*, an archaeon that is distantly related from a phylogenetic perspective but occupies the same ecological niche as coal heaps.

We will discuss lateral gene transfer in Chapter 14.

Draft Sequences of Human Genome (2001)

Two groups published the completion of a draft version of the human genome. This was accomplished by the International Human Genome Sequencing Consortium (IHGS, 2001) and by a consortium led by Celera Genomics (Table 12.11) (Venter et al., 2001). The reports both arrive at the conclusion that there are about 30,000–40,000 protein-coding genes in the genome, an unexpectedly small number. Analysis of the human genome sequence will have vast implications for all aspects of human biology (see Chapter 17).

The bacterial genomes that are sequenced continue to have interesting features. *Mycoplasma pulmonis* has one of the lowest guanine–cytosine (GC) contents that have

TABLE 12-10 Genome-Sequencing Projects Completed in 2000

Organism	Size	Accession	Reference
Halobacterium sp. NRC-1 (A)	2,014,239 bp	NC_002607	Ng et al., 2000
Thermoplasma acidophilum (A)	1,564,906 bp	NC_002578	Ruepp et al., 2000
Bacillus halodurans (B)	4,202,353 bp	BA000004	Takami et al., 2000
Buchnera sp. APS (B)	640,681 bp	NC_002528	
Campylobacter jejuni (B)	1,641,481 bp	NC_002163	Parkhill et al., 2000a
Chlamydia muridarum (B)	1,069,412 bp 1,228,267 bp	NC_002178 NC_002491	Read et al., 2000
Chlamydia pneumoniae AR39 (B)	1,229,853 bp	NC_002179	Read et al., 2000
Chlamydia trachomatis (B)	1,069,412 bp	NC_000117	Read et al., 2000; see Stephens et al., 1998
Neisseria meningitidis MC58 (B)	2,272,351 bp	NC_002183	Tettelin et al., 2000
Neisseria meningitidis Z2491 (B)	2,184,406 bp	NC_002203	Parkhill et al., 2000b
Pseudomonas aeruginosa (B)	6,264,403 bp	NC_002516	Stover et al., 2000
Ureaplasma urealyticum (B)	751,719 bp	NC_002162	Glass et al., 2000
Vibrio cholerae (B)	2,961,149 bp 1,072,315 bp	NC_002505 NC_002506	Heidelberg et al., 2000 Heidelberg et al., 2000
Xylella fastidiosa (B)	2,679,306 bp	NC_002488	Simpson et al., 2000
Drosophila melanogaster (E)	137 Mb	NC_004354 NT_003779 NT_003778 NT_037436 NT_033777 NC_004353	Adams et al., 2000
Arabidopsis thaliana (E)	125 Mb	NC_003070 NC_003071 NC_003074 NC_003075 NC_003076	The Arabidopsis Genome Initiative, 2000
Homo sapiens chromosome 21 (E)	33.8 Mb	NT_002836 NT_002835 NT_003545 NT_001715 NT_001035	Hattori et al., 2000

Abbreviations: A, archaeon; B, bacterium; E, eukaryote.

been described, 26.6% (Chambaud et al., 2001). The genome of *Mycobacterium leprae*, the bacterium that causes leprosy, has undergone massive gene decay, with only half the genome coding for genes (Cole et al., 2001). Analysis of the *Pasterurella multocida* genome suggests that the radiation of the γ subdivision of proteobacteria, which includes *H. influenzae* and *E. coli* and other pathogenic gram-negative bacteria, occurred about 680 MYA (May et al., 2001). The *Sinorhizobium meliloti* genome consists of a circular chromosome and two additional megaplasmids (Galibert et al., 2001). Together, these three elements total 6.7 Mb, expanding our view of the diversity of bacterial genome organization.

Cryptomonads are a type of algae that contain one distinct eukaryotic cell (a red alga, with a nucleus) nested inside another cell (see Box 16.2). This unique arrangement derives from an ancient evolutionary fusion of two organisms. That red algal nucleus, termed a nucleomorph, is the most gene-dense eukaryotic genome known.

TABLE 12-11 Genome-Sequencing Projects Completed in 2001

Organism	Size	Accession	Reference
Pyrococcus abyssi (A)	1,765,118 bp	NC_000868	R. Heilig, 2001
Sulfolobus solfataricus (A)	2,992,245 bp	NC_002754	She et al., 2001
Sulfolobus tokodaii (A)	2,694,765 bp	NC_003106	Kawarabayasi et al., 2001
Caulobacter crescentus (B)	4,016,942 bp	NC_002696	Nierman et al., 2001
Escherichia coli 0157:H7 (B)	5,498,450 bp	NC_002695	Perna et al., 2001; Hayashi et al., 2001; see: Blattner et al., 1997
	5,528,970 bp	AE005174	Perna et al., 2001
Mycobacterium leprae (B)	3,268,203 bp	NC_002677	Cole et al., 2001
Mycoplasma pulmonis (B)	963,879 bp	NC_002771	Chambaude et al., 2001
Pasterurella multocida (B)	2,257,487 bp	AE004439	May et al., 2001
Sinorhizobium meliloti (B)	6.7 Mb	NC_003047	Galibert et al., 2001
Streptococcus pneumoniae (B)	2,160,837	AE005672	Tettelin et al., 2001
Streptococcus pyogenes (B)	1,852,442 bp	AE004092	Ferretti et al., 2001
Encephalitozoon cuniculi (E)	2.5 Mb	AL391737 and AL590442 to AL590451	Katinka et al., 2001
Guillardia theta nucleomorph genome (E)	551,264 bp	NC_002751	Douglas et al., 2001
Homo sapiens (E)	3,300 Mb	Various	IHGSC, 2001; Venter et al., 2001

Abbreviations: A, archeon; B, bacterium; E, eukaryote.

Its genome was sequenced (Douglas et al., 2001) and found to be dense (1 gene per 977 bp) with ultrashort noncoding regions.

Continuing Rise in Completed Genomes (2002)

In the year 2002, dozens more microbial genomes were sequenced. Of the eukaryotes (Table 12.12), the fission yeast *Schizosaccharomyces pombe* was found to have the smallest number of protein-coding genes (4824) (Wood et al., 2002). The genomes of both the malaria parasite *Plasmodium falciparum* and its host, the mosquito *Anopheles gambiae,* were reported (Holt et al., 2002). Additionally, the genome of the rodent malaria parasite *Plasmodium yoelii yoelii* was determined and compared to that of *P. falciparum* (Carlton et al., 2002). These projects will be described in Chapter 16.

TABLE 12-12 Eukaryotic Genome-Sequencing Projects Completed in 2002

Organism	Size	Accession	Reference
Anopheles gambiae (E)	278Mb	AAAB00000000	Holt et al., 2002
Plasmodium falciparum (E)	22.8 Mb	NC_002375	Gardner et al., 2002
Plasmodium yoelii yoelii (E)	23.1 Mb	AABL00000000	Carlton et al., 2002
Schizosaccharomyces pombe (E)	13.8 Mb	NC_003424 NC_003423 NC_003421	Wood et al., 2002

Abbreviation: E, eukaryote.

Overview of Genome Analysis

We have surveyed completed genome projects from a chronological point of view. There are many questions associated with genome sequencing. Which genomes are sequenced? How is it accomplished? How big are genomes? When genomic DNA is sequenced, what are its main features (e.g., genes, regulatory regions, repetitive elements) and how are they determined? Even the goals of sequence analysis are evolving as we learn what questions to ask and what tools are available to address those questions.

We will next examine the process of sequencing a genome in a number of distinct phases, from the selection of an appropriate genome to sequencing the DNA to genome annotation (Fig. 12.11). Computational genomics has been reviewed by Stein (2001) and Koonin (2001).

Selection of Genome for Sequencing

The choice of which genome to sequence depends on several main factors. The selection criteria change over time as technological advances reduce costs and as genome-sequencing centers gain experience in this new endeavor:

Genome Size.

- For a microbial genome, the size is typically several megabases (millions of base pairs), and a single center often has the resources to complete the entire project. For larger (typically eukaryotic) genomes, international collaborations are often established to share the effort (see Chapter 16). For the

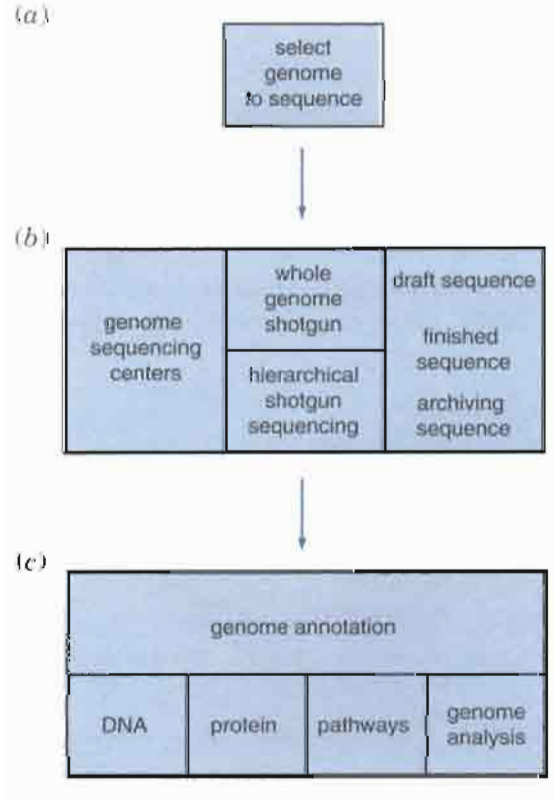


FIGURE 12.11. Overview of the process of sequencing and analyzing a genome. (a) *The selection of which genome to sequence involves decisions about cost, relevance to biological principles, and relevance to disease.* (b) *A variety of genome-sequencing centers perform genome sequencing by approaches such as whole-genome shotgun (WGS) sequencing, hierarchical WGS, or both. This DNA sequencing is performed in stages, including draft and finished sequencing, and the results are archived.* (c) *The completed sequence is annotated at the level of DNA (e.g., to identify repetitive elements, nucleotide composition, and protein-coding genes), at the level of predicted proteins, and at the level of predicted cellular pathways. Additionally, a variety of genomewide analyses may be performed, such as comparisons between genomes or phylogenomics (see below).*

Human Genome Project (Chapter 17), about 1000 bp have been sequenced per second, 24 hours a day, in the past several years.

Cost.

- The total worldwide cost of producing a draft sequence of the human genome by a public consortium was about $300 million. Some of the major sources for funding of genome sequencing have been the National Institutes of Health (Bethesda, Maryland) and the Wellcome Trust (England).

As of June 2003, the Wellcome Trust Sanger Institute (▸http://www.wellcome.ac.uk/) has supported the sequencing of over 2 Gb of DNA from several dozen organisms (▸http://www.sanger.ac.uk/Info/Statistics/).

Relevance to Human Disease.

- For example, by sequencing the chimpanzee genome, we may learn why these animals are not susceptible to diseases that afflict humans, such as malaria and AIDS (Chapter 18).

Relevance to Basic Biological Questions.

- For example, the chicken provides a nonmammalian vertebrate system that is widely used in the study of development. The analysis of protozoan genomes can illustrate the evolutionary history of the eukaryotes.

Relevance to Agriculture.

- The chicken, cow, and honeybee genome sequences are expected to benefit agriculture in a variety of ways, such as leading to strategies to protect these organisms from disease.

The National Human Genome Research Institute (NHGRI) is an example of a federal agency that sponsors genome-sequencing projects. The NHGRI solicits proposals (called "white papers") in which research groups request funding to complete a genome-sequencing project and provide the rationale for selecting that particular organism (Table 12.13). You can read these proposals at the NHGRI website.

Should an Individual from a Species, Several Individuals, or Many Individuals be Sequenced?

Ultimately, it will be important to determine the entire genomic sequence from multiple individuals of a species in order to correlate the genotype with the

TABLE 12-13 Proposed Genomes for Sequencing, 2002

Organism	Species	Priority
Chicken	*Gallus gallus*	High
Chimpanzee	*Pan troglodytes*	High
Cow	*Bos taurus*	High
Dog	*Canis familiaris*	High
Fungi	15 species	High
Honey bee	*Apis mellifera*	High
Oxytricha (protozoan ciliate)	*Oxytricha trifallax*	High
Sea urchin	*Strongylocentrotus purpuratus*	High
Tetrahymena	*Tetrahymena thermophila*	High
Trichoplax (lower metazoan)	*Trichoplax adhaerens*	Moderate
Rhesus macaque	*Macaca mulatta*	Moderate

Source: Adapted from the National Human Genome Research Institute website at the NIH, ▸http://www.genome.gov/page.cfm?pageID=10002154. This site provides links to "white paper" proposals from research groups proposing each sequencing project.

phenotype. In the case of humans (Chapter 17), the public consortium's Human Genome Project initially involved the sequencing and analysis of genomic DNA from individuals from many ethnic backgrounds, both male and female.

For viruses such as human immunodeficiency virus (HIV-1 and HIV-2), the virus rapidly undergoes enormous numbers of DNA changes, making it necessary to sequence many thousands of independent isolates (Chapter 13). This is practical to achieve because the genome is extremely small (<10 kb). In some cases, comparison of different bacterial strains reveals why one is harmless to humans while another is highly pathogenic. (Table 14.10) Such a comparison has been performed for a strain of *E. coli* that normally inhabits the human gut and another strain that causes severe, sometimes fatal disease (Chapter 14).

How Big Are Genomes?

A graphical overview of the sizes of various genomes is presented in Figure 12.12. Viral genomes range from 1 to 350 kb (Chapter 13). In haploid genomes such as bacteria (Chapter 14), the genome size (or C value) is the total amount of DNA in the genome. In diploid or polyploid organisms, the genome size is the amount of DNA in the unreplicated haploid genome (such as the sperm cell nucleus). Bacterial genomes vary in size over about a 22-fold range from about 500,000 bp

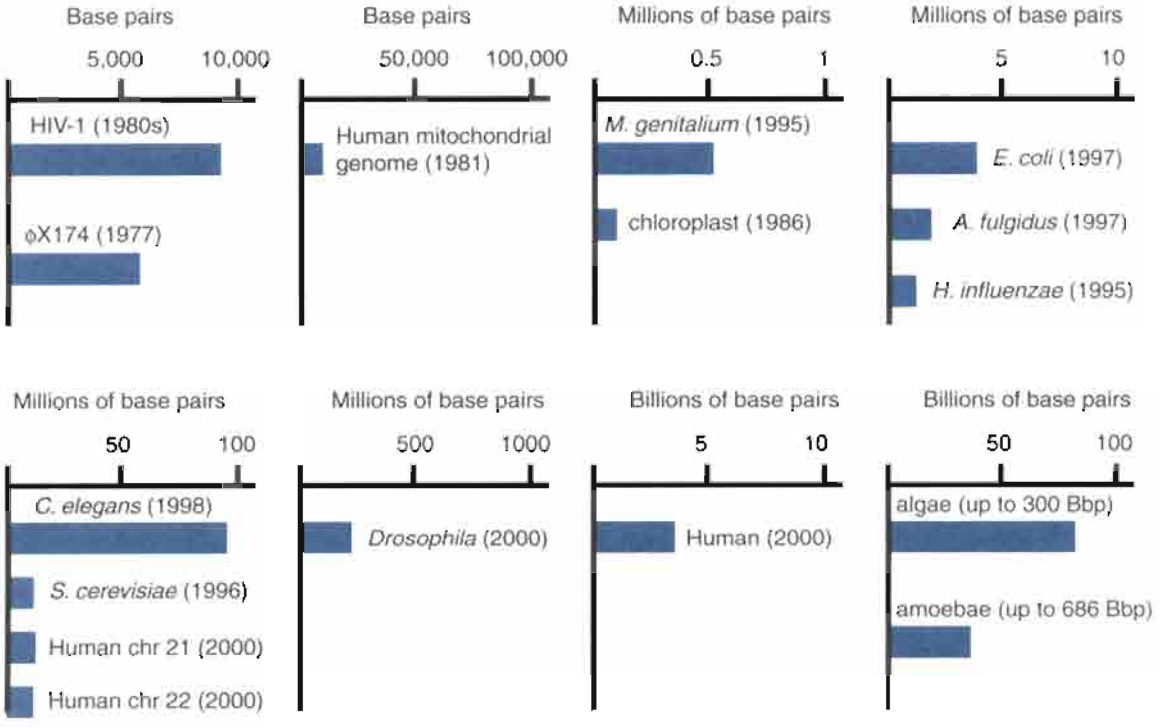


*FIGURE 12.12. Comparison of the sizes of various genomes. The x axis on each graph represents a 10-fold change in scale. For bacterial genomes, the genome size ranges from a mere 580,000 bp (*M. genitalium, *with 470 protein-coding genes is the smallest sequenced genome) to cyanobacteria with genome sizes of 13 Mb. This is a 22-fold range. For eukaryotic genomes, the range is from the 8 Mb of some fungi to 686 Gb for some amoebae. This range is over 75,000-fold and has been called the C value paradox (see Chapter 16). The C value is the total amount of DNA in the genome, and the paradox is the relation between complexity of a eukaryote and its amount of genomic DNA.*

(*M. genitalium*) to 13 Mb (currently the largest sequenced prokaryotic genome, *Streptomyces coelicolor*, is 8.7 Mb).

Among eukaryotes, there is about a 75,000-fold range in genome sizes from 8 Mb for some fungi to 686 Gb (gigabases) for some amoebae. The so-called *C* value paradox is that some organisms with extremely large *C* values are morphologically simple and appear to have a modest number of protein-coding genes. We will explore this paradox in Chapter 16.

Genome-Sequencing Centers

Large-scale sequencing projects are conducted at centers around the world. Twenty sequencing centers contributed to the production of a draft version of the human genome (2001) (Table 17.6). These centers were also supported by the NIH and the EBI. All of these centers have also been involved in sequencing the genomes of other organisms as well.

In addition to the centers listed in Table 17.6, there are many other important genome-sequencing centers. These include The Institute for Genomic Research (TIGR), which has contributed many dozens of prokaryotic and eukaryotic genome-sequencing projects, and Celera Genomics, which produced a draft version of the human genome in 2001.

Lists of genome-sequencing centers are offered at the NCBI (▸ http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/links.html), at the UK Human Genome Mapping Project Resource Centre (▸ http://www.hgmp.mrc.ac.uk/GenomeWeb/genome-centres.html), and ▸ http://igweb.integratedgenomics.com/GOLD/Genomiclinks.html.

TIGR's website is at ▸ http://www.tigr.org/ and Celera's is at ▸ http://www.celera.com/. Both are located in Rockville, Maryland.

Sequencing Genomes: Strategies

There are two main approaches to sequencing genomes. The first is whole-genome shotgun (WGS) sequencing. Fred Sanger first applied this approach in the sequencing of bacteriophage ϕX174: Randomly selected fragments of genomic DNA were isolated, sequenced, and then assembled to derive a complete sequence. The application of this approach to an entire organismal genome was pioneered by Hamilton O. Smith of Johns Hopkins and J. Craig Venter of TIGR (subsequently of Celera Genomics).

The whole-genome shotgun method has been used successfully for most small genomes (i.e., viruses, bacteria and archaea, and eukaryotic genomes that lack large portions of repetitive DNA). Genomic DNA is isolated from an organism and mechanically sheared (or digested with restriction enzymes). The fragments are subcloned into small-insert libraries (e.g., 2-kb fragments), and large-insert libraries (e.g., 10–20 kb). Clones are sequenced from both ends (i.e., both "top" strand and "bottom" strand), and then the sequences are assembled. A typical sequencing reaction generates about 500 bp of sequence data. These small amounts of sequence are assembled into contiguous transcripts ("contigs") and then into a map of the complete genome. This strategy was employed in the majority of the bacterial and archaeal sequencing projects described in Tables 12.5–12.12. The *H. influenzae* genome was sequenced by whole-genome shotgunning in 1995 (Table 12.5).

A second, related approach is "hierarchical shotgun sequencing" (Fig. 12.13). Genomic DNA is digested and subcloned into BAC libraries. These libraries contain large inserts (100–500 kb). Alternatively, smaller cosmid libraries (with insert sizes of about 50 kb) or plasmid libraries (2–10-kb inserts) are generated. Unlike WGS sequencing, this hierarchical strategy employs clones (contigs) that are mapped definitively to known chromosomal locations. Thus, sequence assembly is focused on a small region of the genome. This approach has been taken for most large,

See Table 12.14 for terminology associated with genome sequencing.

TABLE 12-14 Terminology Used in Genome-Sequencing Projects

Term	Definition
BAC end sequence	The ends of a bacterial artificial chromosome (BAC) have been sequenced and submitted to GenBank; the internal BAC sequence may not be available. When both end sequences from the same BAC are available, this information can be used to order contigs into scaffolds.
Contig	A set of overlapping clones or sequences from which a sequence can be obtained. NCBI contig records represent contiguous sequences constructed from many clone sequences. These records may include draft and finished sequences and may contain sequence gaps (within a clone) or gaps between clones when the gap is spanned by another clone which is not sequenced.
Draft sequence	At least three- to four fold of the estimated clone insert is covered in Phred Q20 bases in the shotgun sequencing stage, as defined for the human genome-sequencing project. Note that the exact definition of "draft" may be different for other genome projects. Clone sequence may contain several pieces of the sequence separated by gaps. The true order and orientation of these pieces may not be known.
Finished sequence	The clone insert is contiguously sequenced with a high-quality standard of error rate of 0.01%. There are usually no gaps in the sequence.
Fragment	A contiguous stretch of a sequence within a clone sequence that does not contain a gap, vector, or other contaminating sequence.
Meld	When two or more fragments overlap in the entire alignable region, these sequences are merged together to make a single longer sequence.
Order and orientation	Sequence overlap information is used to order and orient (ONO) fragments within a large clone sequence.
Scaffold	Ordered set of contigs placed on the chromosome.

Source: Adapted from ▸http://www.ncbi.nlm.nih.gov/genome/guide/build.html.

Regions of heterochromatin contain large segments of highly repetitive DNA (Chapter 16), and cannot be effectively sequenced using WGS or hierarchical approaches. Skaletsky et al. (2003) applied an alternative technique of iterative mapping and sequencing to determine the repetitive sequence of the human Y chromosome.

eukaryotic genomes, including the public consortium's version of the Human Genome Project.

The WGS approach requires the computationally difficult task of fitting contigs together, regardless of which chromosomal region they are derived from. It was thought by some that this approach could not be practically applied to large eukaryotic genomes. However, it was applied to the 120-Mb *D. melanogaster* genome (Adams et al., 2000) in combination with a hierarchical approach and to the human genome (Weber and Myers, 1997; Venter et al., 2001) (see Chapter 17).

Genomic Sequence Data: From Unfinished to Finished

Phred and Phrap (see below) are available at ▸http://www.phrap.org/. They operate on UNIX-based systems.

Raw DNA sequence data are deposited in databases such as the Trace Archives at NCBI and EBI (see below). The raw DNA data are typically read by software such as Phred. This program interprets which bases are sequenced by a DNA sequencing machine and further estimates the quality of each read. Phred then writes the sequences in a format such as FASTA (Fig. 2.10) for further analysis.

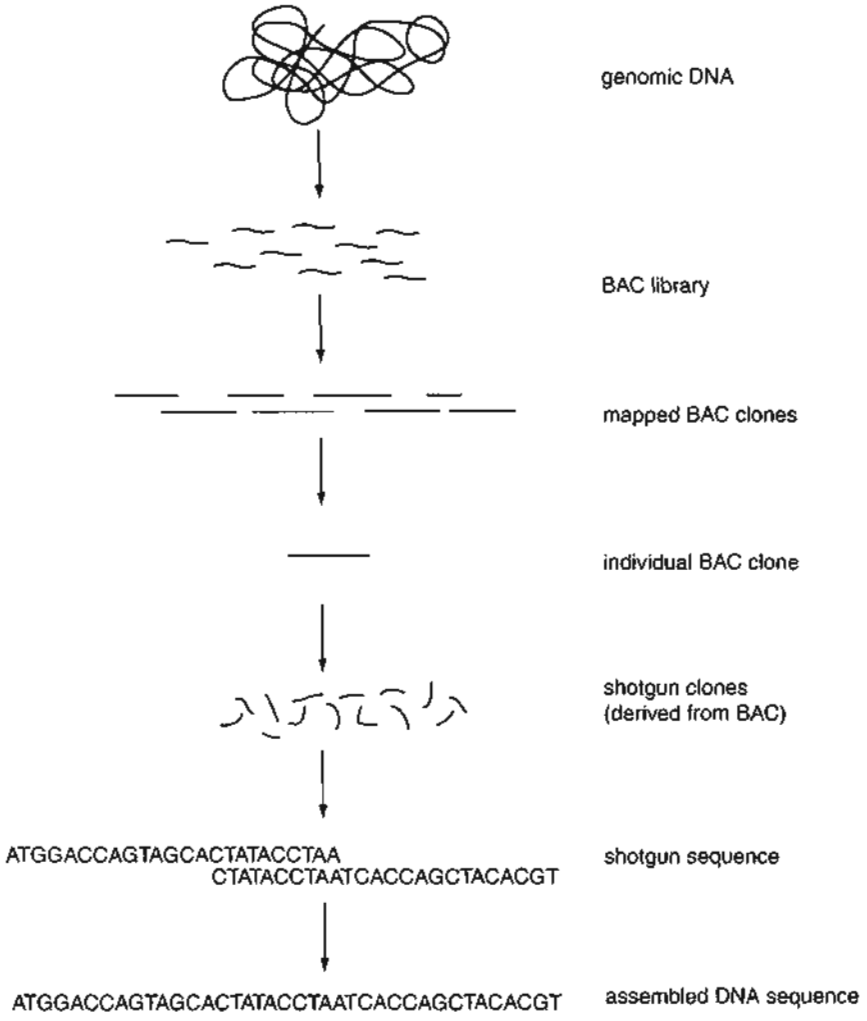


FIGURE 12.13. *Schematic of the hierarchical shotgun sequencing strategy. Genomic DNA is isolated from an organism of interest, fragmented, and inserted into a BAC library. Each BAC clone is 100–500 Kb. BACs are ordered (mapped). Individual BAC clones are fragmented into smaller cDNA clones and sequenced. Individual sequencing reactions are typically 300–700 nucleotides. These "shotgun sequences" are assembled. This process is further illustrated in Figure 17.14. Modified from IHGSC (2001, p. 863). Used with permission.*

At NCBI, raw genomic DNA data are made available through the high-throughput genomic (HTG) sequence division. Accession numbers are assigned to each entry. The HTG database contains sequence data in four phases (Table 12.15). Phase 0 data are typically sequences derived from a single cosmid or BAC. They are likely to have sequencing errors and gaps of indeterminate size. However, the data may still have tremendous usefulness to the scientific community even in this form. For example, if you are performing BLAST searches and are looking for novel homologs to your query, the HTG division may contain useful information. Phase 1 data may consist of sequencing reads from contigs derived from a larger clone (e.g., a

TABLE 12-15 High-Throughput Genomic Records at GenBank Defined in Four Phases

Status	Location	Definition
Phase 0	HTG division	Single-few pass reads of a single clone (not contigs)
Phase 1	HTG division	Unfinished, may be unordered, unoriented contigs, with gaps
Phase 2	HTG division	Unfinished, ordered, oriented contigs, with or without gaps
Phase 3	Primary division	Finished, no gaps (with or without annotations)

Source: From ▸http://www.ncbi.nlm.nih.gov/HTGS/.

BAC clone) in which the order of the contigs is unknown and their orientation (top strand or bottom strand) is also unknown. The sequence is defined as unfinished, and it still contains gaps.

In the finished state (phase 2), the contigs are ordered and oriented properly, and the error rate must be 10^{-4} or less.

For examples of phases 1, 2, and 3 sequences in GenBank, see ▸ http://www.ncbi.nlm.nih.gov/HTGS/examples.html.

The assembly process involves the collection of individual sequences (phase 0), the closing of gaps, and the lowering of the error rate. This process can be performed using a variety of software packages, such as Phrap (and its graphical viewer, Consed), Assembler, and Sequencher. For either the whole-genome sequencing or the hierarchical approach, after the shotgun phase is complete, the next step is to assemble contigs. This is accomplished in a process called finishing. The goal of finishing is to identify gaps in the tile path and to close them. Ideally, this process results in a single contiguous DNA sequence that spans all the contigs. The finishing process can be performed manually by experts or in an automated fashion with a program such as Autofinish (Gordon et al., 2001).

Genomic sequencing projects often rely heavily on expressed sequence tags (ESTs) to help define the protein-coding genes. Transcripts that are expressed (i.e., RNA molecules) are converted to cDNA, incorporated into libraries, and sequenced. Such cDNAs are ESTs. While they do not reveal some information about the corresponding genomic DNA such as the sequence of introns, they are invaluable in identifying expressed genes (see Figure 12.17 below).

When Has a Genome Been Fully Sequenced?

Typically, a genome is sequenced with 5- to 10-fold coverage to maximize the likelihood that it has been completely sequenced. The greatest technical challenge is to resolve the sequences of the long regions of repetitive DNA found in eukaryotic genomes (see below). To date, the human genome is the only one for which large regions of repetitive DNA sequence are being carefully sequenced.

It is possible to estimate the amount of DNA that is sequenced as a function of fold coverage (Table 12.16). The probability a base is not sequenced was derived by Lander and Waterman (1988) and is given by

$$P_0 = e^{-c} \tag{12.1}$$

where c is the fold coverage and is given by

$$c = \frac{LN}{G} \tag{12.2}$$

and where LN is the number of bases sequenced, L being the read length and N the number of reads, and e is the constant 2.718. These results show that to achieve an error rate of 1 in 10,000 (0.01%), it is theoretically necessary to obtain nine-fold coverage of the genome. With fivefold coverage, an error rate of 0.6% is expected.

Some regions of genomic DNA yield ambiguous sequencing results. This can occur in regions of extremely high or low GC content or unusual secondary structure. These areas are routinely resolved ("finished") by directed sequencing of the region in question with specific oligonucleotide primers (Tettelin et al., 1999). The sequencing of the malaria parasite *P. falciparum* was especially difficult to achieve because adenine and thymine comprise about 80% of the genome (Gardner et al., 2002). For this reason, a chromosome-based approach was needed to sequence this 23-Mb genome (Chapter 16).

TABLE 12-16 Probability That a Base Is Sequenced According To Equation 12.1

Fold Coverage	P_0	Percent Not Sequenced	Percent Sequenced
0.25	$e^{-0.25} = 0.78$	78	22
0.5	$e^{-0.5} = 0.61$	61	39
0.75	$e^{-0.75} = 0.47$	47	53
1	$e^{-1} = 0.37$	37	63
2	$e^{-2} = 0.135$	13.5	87.5
3	$e^{-3} = 0.05$	5	95
4	$e^{-4} = 0.018$	1.8	98.2
5	$e^{-5} = 0.0067$	0.6	99.4
6	$e^{-6} = 0.0025$	0.25	99.75
7	$e^{-7} = 0.0009$	0.09	99.91
8	$e^{-8} = 0.0003$	0.03	99.97
9	$e^{-9} = 0.0001$	0.01	99.99
10	$e^{-10} = 0.000045$	0.005	99.995

Source: Adapted from ▸ http://www.genome.ou.edu/poisson.calc.html and Lander and Waterman (1988).

Repository for Genome Sequence Data

Raw sequence data for the genome-sequencing projects of several organisms have been deposited in the Trace Archive located at both NCBI and EBI. All entries in this archive are given a Trace Identifier (Ti) number. The archive can be searched by several criteria (such as query by Ti or sequencing center or by BLAST).

Search the mouse trace archive with a mouse RBP clone (M74527) and the output contains several Ti matches (Fig. 12.14*a*). By clicking on the link to a Ti record, the sequence data can be obtained in the FASTA format or as a trace of the dye termination reaction used to sequence the DNA (Fig. 12.14*b*).

A specialized Trace Archive BLAST server is available at ▸ http://www.ncbi.nlm.nih.gov/blast/mmtrace.html.

Table 12.17 summarizes several principal organisms for which trace archive data are available.

Genome Annotation: Features of Genomic DNA

When a genome is sequenced, we learn its exact size and we obtain the complete (or nearly complete) nucleotide sequence. Genome annotation is the process by which the landscape of genomic DNA is surveyed, and key features of the DNA are described (Stein, 2001). An example of an automated pipeline for genome annotation from the Ensembl server is outlined in Figure 12.15.

Three fundamental questions may be asked about the nature of this sequence:

1. What is the overall GC content or other nucleotide composition? Many eukaryotic genomes are characterized by a GC content of about 35–45%, while bacteria display a far wider range (Fig. 12.16).
2. What are the repetitive DNA sequences and where are they? Programs such as RepeatMasker can identify and mask repetitive elements such as Alu repeats. We will discuss these repeats in Chapters 15–17. Programs such as GLIMMER and GRAIL (see below) incorporate algorithms that identify repetitive elements in genomic DNA.

We will show examples of repetitive DNA, and the software used to identify and mask it, in Figures 16.4–16.10.

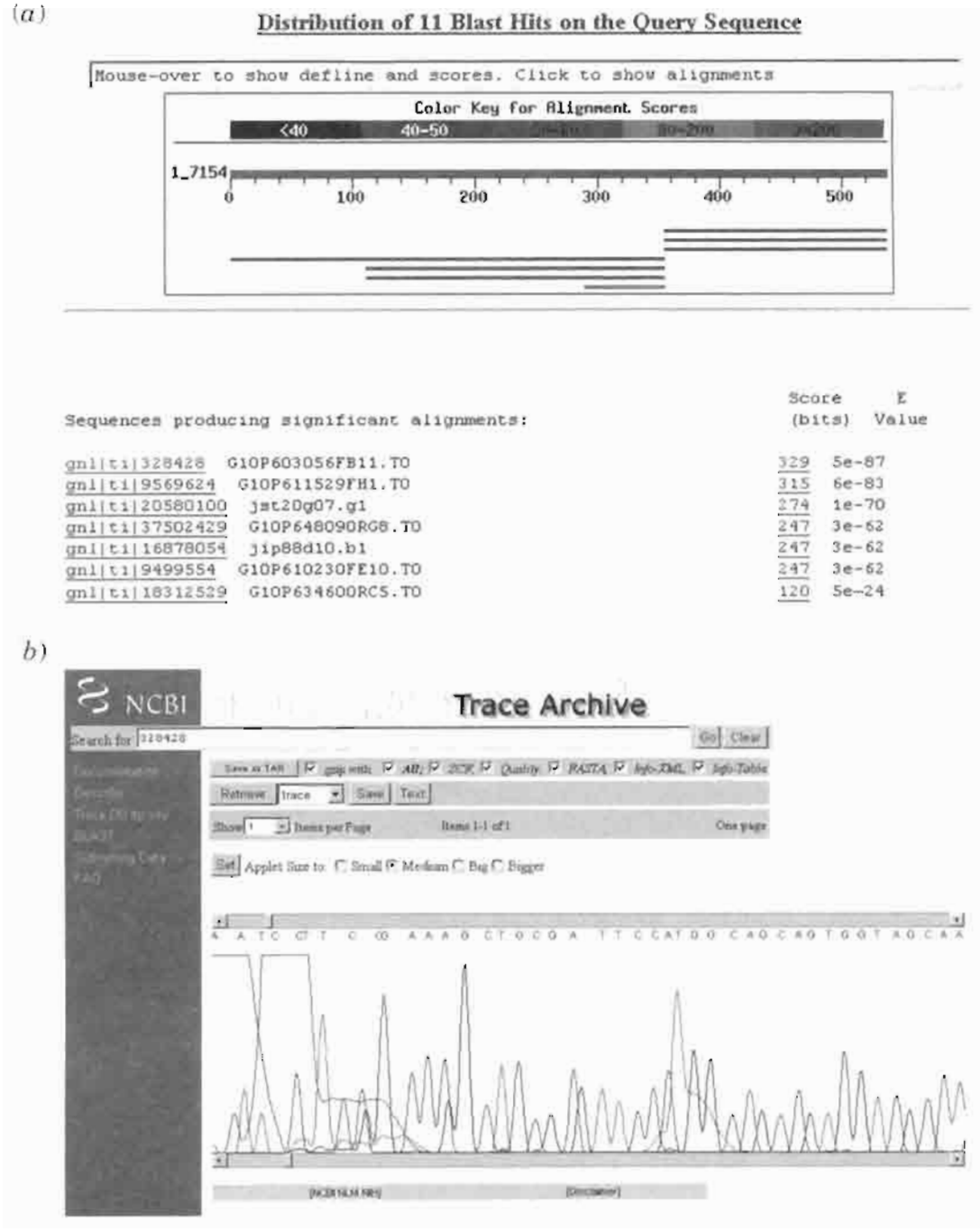


FIGURE 12.14. *The trace archive is a repository of raw data from genome-sequencing projects. It is accessed from the front page of NCBI (▶ http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?) or from the Trace Server at EBI (▶ http://trace.ensembl.org/).* (a) *A BLASTN search with mouse* RBP4 *as a query results in 11 trace archive matches. The pattern of hits (several to the 3′ end of the query and several others to the 5′ end) occurs because the query is an expressed transcript, and there is an intron corresponding to position 350.* (b) *Clicking on the first database match (ti 328428) provides access to the sequence in FASTA format or the trace from the DNA-sequencing reaction. This trace allows you to evaluate the quality of the raw data that underlie a genomic DNA record. In some cases, the dye termination reaction (or alternate DNA-sequencing technology) yields ambiguous results, and access to the raw data allows the user to make an informed decision about the quality of the sequence call.*

TABLE 12-17 Current Contents of Trace Archives of Sequence Data for Selected Organisms Listed Alphabetically (March 2003). The Trace Archives Contain Data from WGS, EST, and BAC-based Projects.

Organism	Common Name	Number of Traces
Anopheles gambiae	Mosquito	4,888,159
Bos taurus	Cow	426,683
Caenorhabditis briggsae	Worm	2,354,917
Danio rerio	Zebrafish	12,125,895
Felis catus	Cat	460,012
Glycine max	Soybean	157,399
Homo sapiens	Human	18,241,156
Mus musculus	Mouse	57,591,595
Rattus norvegicus	Rat	37,781,311
Takifugu rubripes	Pufferfish	1,953,192
Tetraodon nigroviridis	Pufferfish	2,688,931
Zea mays	Corn	69,888

Source: From ▸http://trace.ensembl.org/ or ▸http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?.

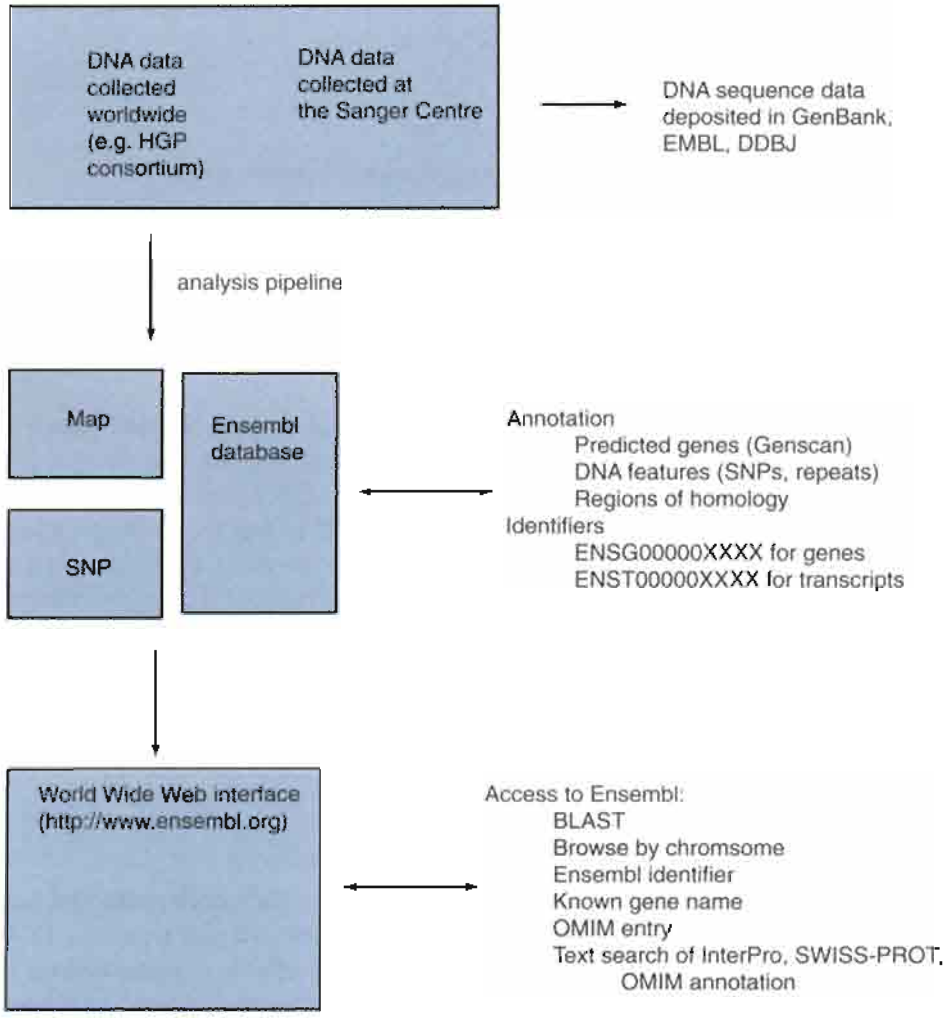


FIGURE 12.15. *Overview of the Ensembl annotation pipeline (from ▸http://www.ensembl.org/Docs/enstour/). Ensembl is a joint EBI-EMBL and Sanger Institute project that automatically tracks and annotates DNA sequence data from the Human Genome Project and other sequencing projects (e.g., mouse, rat zebrafish, fugu, mosquito, fruitfly, and two nematodes).*

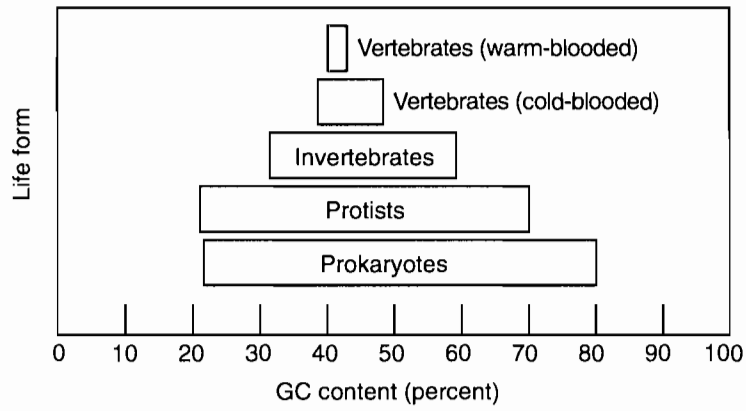


FIGURE 12.16. *Guanine plus cytosine (GC) content of prokaryotes, protists, invertebrates, and vertebrates. Note that most eukaryotic genomes have 40–45% G+C content, while bacteria and archaea have a far wider range. This figure is adapted from Bernardi and Bernardi (1990) based on studies in the 1970s and 1980s. Recent eukaryotic genome sequencing projects (described in Chapter 16) reveal that GC content for various organisms includes 19.4%* (P. falciparum), *22.2% (the slime mold* Dictyostelium discoideum), *34.9%* (A. thaliana), *36%* (C. elegans), *38.3%* (S. cerevisiae), *41.1% (human), 42%* (M. musculus), *and 43.3%* (O. sativa). *For sequenced prokaryotes, GC content values range from 26%* (Ureaplasma urealyticum parvum) *to 72%* (Streptomyces coelicolor) *(Table 14.3). Used with permission.*

3. How many genes (protein-coding sequences) are present? Genes may be identified by a number of features, including:
   - Gene-specific codon bias
   - Absence of repetitive DNA sequences
   - Presence of signals such as promoter region-specific motifs

These features of genomic DNA are substantially different between prokaryotes and eukaryotes. We will thus consider them in more detail in Chapters 14 (on bacteria and archaea) and 15–17 (on eukaryotes).

Annotation of Genes in Bacteria

Bacterial and archaeal genomes have both genes and additional, relatively small intergenic regions. Typically, these genomes are circular, and there is about one gene in each kilobase of genomic DNA. For prokaryotes, genes are most simply identified by the presence of long open reading frames (ORFs) that are greater in length than some cutoff value such as 90 nucleotides (30 amino acids; a protein of about 3 kD). Programs such as GLIMMER and GenMark efficiently locate genes in bacterial genomic sequence (reviewed in Baytaluk et al., 2002) (Table 12.18).

Glimmer is a program for the identification of genes in prokaryotic DNA. The program requires two inputs: a genomic DNA sequence file (in FASTA format) and a set of Markov models for genes. We will examine a sample GLIMMER output (see Fig. 14.7).

Annotation of Genes in Eukaryotes

In contrast to bacterial genomes, eukaryotic genomes contain both genes and large amounts of noncoding DNA. This noncoding material includes repetitive DNA, genes that have regulatory functions, and introns that interrupt exons and are removed from mature RNA transcripts. A major focus of genome-sequencing projects is to identify all the genes in a genome. However, it is necessary to define the variety

TABLE 12-18 Gene-Finding Programs as Summarized at Southwest Biotechnology and Bioinformatics Center (▸ http://www.swbic.org/links/1.4.3.2.php)

Additional programs are listed at that website

Resource	Source/Description	URL
BCM Gene Finder	Provides several tools to identify genes	▸http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html
GeneFind	Protein family identification of queries and keyword searches	▸http://pir.georgetown.edu/gfserver
GeneMark	Provides several gene prediction programs	▸http://opal.biology.gatech.edu/GeneMark/
Gene Parser	Identifies protein-coding regions in genomic DNA	▸http://beagle.colorado.edu/~eesnyder/GeneParser.html
GeneScan	Identification of complete gene structures in vertebrate genomic DNA	▸http://genes.mit.edu/GENSCAN.html
GLIMMER	A system for finding genes in microbial DNA	▸http://www.tigr.org/software/glimmer/
GRAIL	Gene Recognition and Assembly Internet Link	▸http://compbio.ornl.gov/Grail-bin/EmptyGrailForm
ORF Finder	Open reading frame finder	▸http://www.ncbi.nlm.nih.gov/gorf/gorf.html
Procrustes	Gene recognition via spliced alignment	▸http://www-hto.usc.edu/software/procrustes/

of genes and the criteria for identifying them. This includes protein-coding genes, pseudogenes, and a variety of RNA genes. We will discuss these in Chapters 15 (on fungi) and 17 (on human), with a particular emphasis on gene finding in Chapter 16 (on eukaryotes). There, we will discuss two principal approaches to gene identification in eukaryotic genomic DNA (Fig. 12.17). The first approach is based on aligning expressed sequences (ESTs or cDNAs) to genomic DNA. Since the ESTs are obtained independently of the genomic DNA sequence, this approach is called "extrinsic."The availability of a full-length cDNA is invaluable in defining the extent of the exons in a gene based on experimental evidence. Alternatively, an "intrinsic" approach is to predict gene structures (exons and introns) solely through analysis of genomic DNA, searching for features such as ORFs, exon/intron boundaries, start and stop codons, and codon usage typical of coding regions. Examples of gene prediction software are presented in Figures 16.12–16.18.

Summary: Questions from Genome-Sequencing Projects

A series of basic questions are associated with virtually all genome sequencing projects:

- Can we identify both protein-coding genes and RNA-coding genes?
- Can we assign a function to these genes?
- Can we determine (or predict) the structure of all the gene products?
- Can we reconstruct transcriptional networks and metabolic signaling pathways associated with each gene product?
- Can we link genotypes to phenotypes? Thus, for example, can we explain why humans vary greatly in their susceptibility to the same disease-causing organism or environmental toxin? Can we explain why two strains of anthrax or herpesvirus vary in their pathogenicity?
- Can we define the evolutionary history of life? This may be accomplished in part through molecular phylogenetic studies and comparative genomics. This approach has been called phylogenomics (Fig. 12.18) (Eisen and Hanawalt, 1999; Eisen and Fraser, 2003).

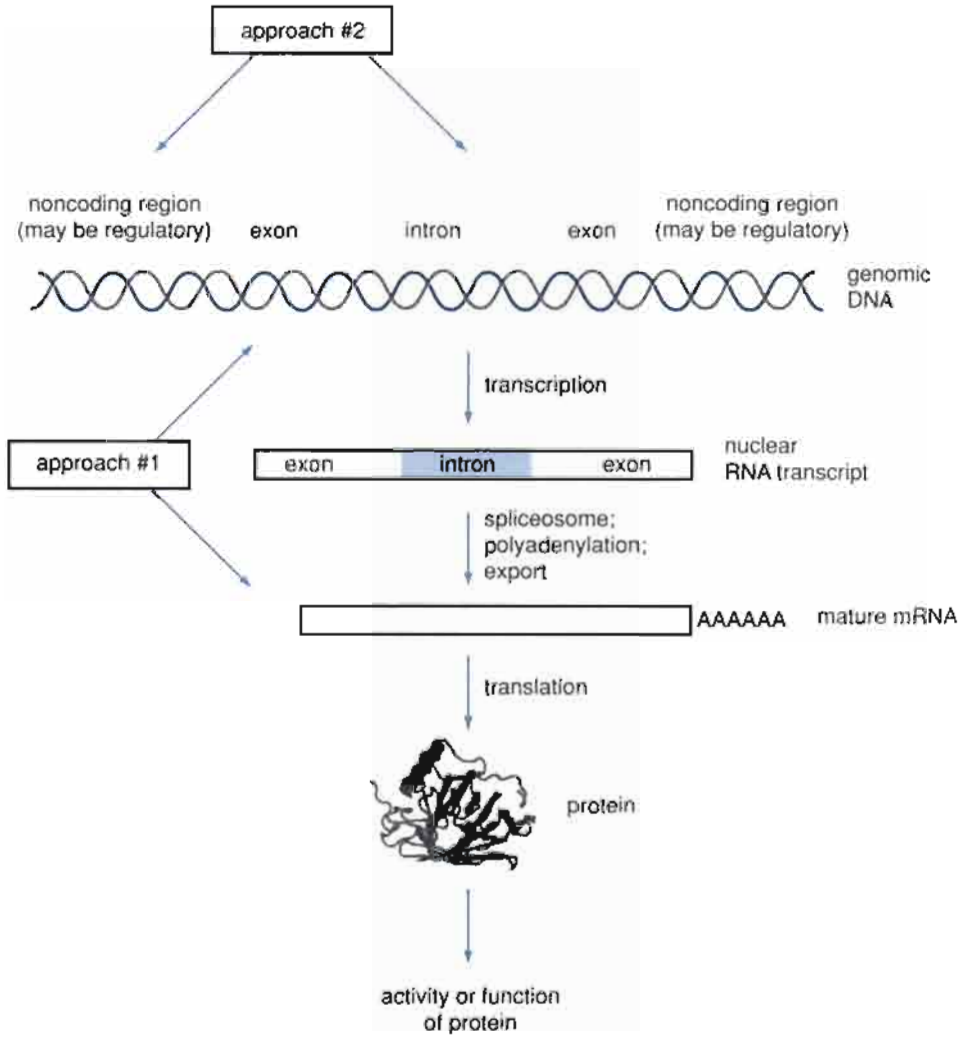


FIGURE 12.17. *There are two principal approaches to finding protein-coding genes in genomic DNA. In the first approach, called homology-based (or extrinsic) gene finding, genomic DNA is compared to ESTs. ESTs are cDNAs that are generated from the RNA of an organism, and thousands to millions of ESTs are available for various organisms. When an EST sequence matches a region of genomic DNA, this provides strong evidence that a protein-coding gene has been identified. In the second approach, algorithm-based (or intrinsic) gene finding, the nucleotide composition of the genomic DNA is analyzed for features such as the presence of a long open reading frame, i.e., a start codon followed by a threshold such as at least 300 nucleotides before a stop codon is encountered. The presence of introns (usually in eukaryotic genomes) complicates this analysis. The base composition of coding regions often differs dramatically from noncoding regions, and this also serves as the basis for gene-finding algorithms using the second approach.*

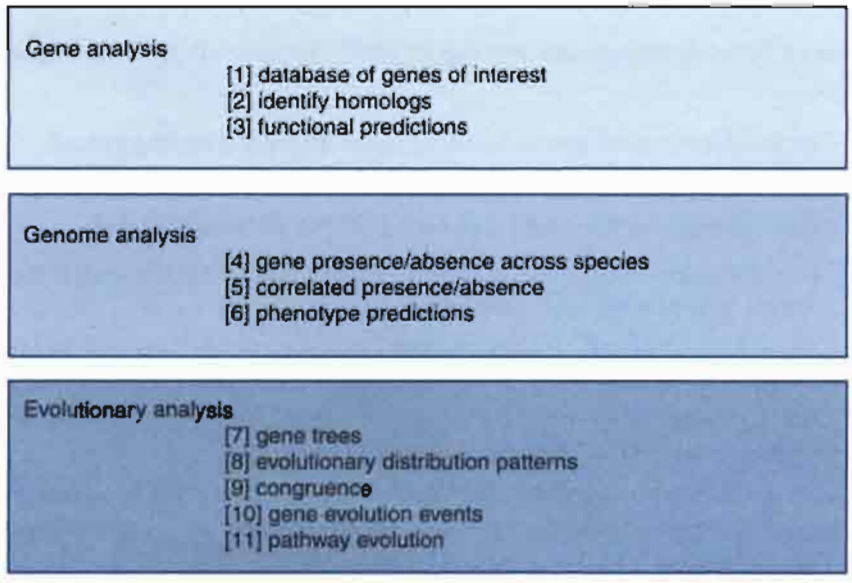


FIGURE 12.18. *Phylogenomics describes an evolution-based approach to genomics. (Adapted from Eisen and Hanawalt, 1999.) Used with permission.*

PERSPECTIVE

Beginning in 1995, we have entered an era in which the completed genome sequence has been determined for many dozens of organisms. Very soon, hundreds of complete genome sequences will become available. With the completion of the human genome sequencing in the year 2003, some call the present state of biology the "postgenomic era."

A major consequence of genome-sequencing projects is that molecular phylogeny has been revolutionized. The present version of the tree of life includes three main branches (bacteria, archaea, and eukaryotes). In the coming years, molecular data will clarify some of the key questions about life on Earth:

- How many species exist on the planet?
- How did life evolve, from 4 BYA up to the present time?
- Why are some organisms pathogenic while close relatives are harmless?
- What mutations cause disease in humans and other organisms?

PITFALLS

While the research community is generating massive amounts of DNA sequence data, there are many pitfalls associated with interpretation of those data. There is an error rate associated with genome sequences (typically one nucleotide per 10,000 in finished DNA). Thus, in evaluating possible polymorphisms or mutations in genomic DNA sequences, it is important to assess the quality of the sequence data. Even if the sequence is correct, algorithms do not yet have complete success in problems such as finding protein-coding genes in eukaryotic DNA; in Chapter 16 we will see examples of genome-sequencing projects (such as rice and human) in which the predicted exons and gene models improve dramatically with each subsequent revision of the genome sequence. (For bacterial DNA, which generally lack introns, the success rate is much higher.) Once protein-coding genes or other types of genes are identified, there are very large numbers of errors in genome annotation (Brenner, 1999). It will be important to carefully assess the basis of functional annotation of genes, and ultimately the problem of gene function must be assessed by biological as well as computational criteria.

DISCUSSION QUESTION

[12-1] If you could decide which genome-sequencing projects to pursue, how would you prioritize the organisms?

PROBLEM

[12-1] Figure 12.1 shows a tree of life based on rRNA sequences. Construct a tree of life based on glyceraldehyde 3-phosphate dehydrogenase (GAPDH) protein sequences. How similar are the trees? What might account for their differences?

SELF-TEST QUIZ

[12-1] The first complete genome to be sequenced was:
(a) *Saccharomyces cerevisiae* chromosome III
(b) *Haemophilus influenzae*
(c) ϕX174
(d) The human mitochondrial genome

[12-2] A typical eukaryotic mitochondrial genome encodes about how many proteins (excluding RNAs)?
(a) 10
(b) 100
(c) 1000
(d) 10,000

[12-3] The term "whole-genome shotgun sequencing" refers to:
(a) A strategy to sequence an entire genome by breaking up DNA and sequencing using oligonucleotide primers that span the genomic DNA
(b) A strategy to sequence an entire genome by breaking up DNA, cloning it into a library, and sequencing using oligonucleotide primers that correspond to known chromosomal locations (contigs)
(c) A strategy to sequence an entire genome by breaking up DNA, cloning it into a library, identifying the chromosomal location of each large library fragment (contig), sequencing small fragments, then reassembling the fragments into a complete map
(d) A strategy to sequence an entire genome by breaking up DNA, cloning it into a library, sequencing small fragments, then reassembling the fragments into a complete map

[12-4] The biggest problem in predicting protein-coding genes from genomic sequences using algorithms is that:
(a) The software is difficult to use.
(b) The false-negative rate is high: Many exons are missed.
(c) The false-positive rate is high: Many exons are falsely assigned.
(d) The false-positive rate is high: Many exons have unknown function.

[12-5] The cost of sequencing genomic DNA at a major genome sequencing center is approximately
(a) one cent per base
(b) one dollar per base
(c) ten dollars per base
(d) $1000 per genome

[12-6] Many hundreds of genomes have now been completely sequenced. In terms of the genus and species of various genomes, the majority of these are
(a) viral
(b) bacterial
(c) archaeal
(d) organellar (mitochondrial and plastid)
(e) eukaryotic

[12-7] There are many criteria that have been applied to decide which genomes to sequence. Which of the following is NOT one of the major criteria?
(a) cost of the project
(b) relevance to human disease
(c) size of the genome
(d) number of genes in the genome
(e) relevance to fundamental biological processes

REFERENCES

Anderson, S., et al. Sequence and organization of the human mitochondrial genome. *Nature* **290,** 457–465 (1981).

Adams, M. D., et al. The genome sequence of *Drosophila melanogaster. Science* **287,** 2185–2195 (2000).

Alm, R. A., et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori. Nature* **397,** 176–180 (1999).

The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* **408,** 796–815 (2000).

Anderson, L. Identification of mitochondrial proteins and some of their precursors in two-dimensional electrophoretic maps of human cells. *Proc. Natl. Acad. Sci. USA* **78,** 2407–2411 (1981).

Andersson, S. G., et al. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396,** 133–140 (1998).

Baytaluk, M. V., Gelfand, M. S., and Mironov, A. A. Exact mapping of prokaryotic gene starts. *Brief Bioinform.* **3,** 181–194 (2002).

Benton, M. J., and Ayala, F. J. Dating the tree of life. *Science* **300,** 1698–1700 (2003).

Bernardi, G., and Bernardi, G. Compositional transitions in the nuclear genomes of cold-blooded vertebrates. *J. Mol. Evol.* **31,** 282–293 (1990).

Blackmore, S. Environment. Biodiversity update—progress in taxonomy. *Science* **298,** 365 (2002).

Blattner, F. R., et al. The complete genome sequence of Escherichia coli K-12. *Science* **277,** 1453–1474 (1997).

Brenner, S. E. Errors in genome annotation. *Trends Genet.* **15,** 132–133 (1999).

Bult, C. J., et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* **273,** 1058–1073 (1996).

Carlton, J. M., et al. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature* **419,** 512–519 (2002).

Casjens, S., et al. A bacterial genome in flux: The twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi*. *Mol. Microbiol.* **35,** 490–516 (2000).

Chambaud, I., et al. The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. *Nucleic Acids Res.* **29,** 2145–2153 (2001).

Cole, S. T., et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393,** 537–544 (1998).

Cole, S. T., et al. Massive gene decay in the leprosy bacillus. *Nature* **409,** 1007–1011 (2001).

Dandekar, T., et al. Re-annotating the *Mycoplasma pneumoniae* genome sequence: Adding value, function and reading frames. *Nucleic Acids Res.* **28,** 3278–3288 (2000).

Deckert, G., et al. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392,** 353–388 (1998).

Douglas, S., et al. The highly reduced genome of an enslaved algal nucleus. *Nature* **410,** 1091–1096 (2001).

Dunham, I., et al. The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).

Eisen, J. A. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. *Curr. Opin. Genet. Dev.* **10,** 606–611 (2000).

Eisen, J. A., and Fraser, C. M. Phylogenomics: intersection of evolution and genomics. *Science* **300,** 1706–1707 (2003).

Eisen, J. A., and Hanawalt, P. C. A phylogenomic study of DNA repair genes, proteins, and processes. *Mutat. Res.* **435,** 171–213 (1999).

Fauron, C., Casper, M., Gao, Y., and Moore, B. The maize mitochondrial genome: Dynamic, yet functional. *Trends Genet.* **11,** 228–235 (1995).

Ferretti, J. J., et al. Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc. Natl. Acad. Sci. USA* **98,** 4658–4663 (2001).

Fleischmann, R. D., et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).

Fox, G. E., et al. The phylogeny of prokaryotes. *Science* **209,** 457–463 (1980).

Fraser, C. M., et al. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270,** 397–403 (1995).

Fraser, C. M., et al. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390,** 580–586 (1997).

Fraser, C. M., et al. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281,** 375–388 (1998).

Galibert, F., et al. The composite genome of the legume symbiont *Sinorhizobium meliloti*. *Science* **293,** 668–672 (2001).

Gardner, M. J., et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419,** 498–511 (2002).

The C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans:* A platform for investigating biology. *Science* **282,** 2012–2018 (1998).

Giaever, G., et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418,** 387–391 (2002).

Glass, J. I., et al. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature* **407,** 757–762 (2000).

Goffeau, A., et al. Life with 6000 genes. *Science* **274,** 546, 563–577 (1996).

Gordon, D., Desmarais, C., and Green, P. Automated finishing with autofinish. *Genome Res.* **11,** 614–625 (2001).

Graur, D., and Li, W.-H. *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, MA, 2000.

Haeckel, E. *The Evolution of Man: A Popular Expositon of the Principal Points of Human Ontogeny and Phylogeny.* D. Appleton and Company, New York, 1879.

Hattori, M., et al. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* **405,** 311–319 (2000).

Hayashi, T., et al. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res* **8,** 11–22 (2001).

Hedges, S. B., et al. Genomic timescale for the origin of eukaryotes. *BMC Evol. Biol.* **1,** 1–10 (2001).

Heidelberg, J. F., et al. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406,** 477–483 (2000).

Heilig, R. *Pyrococcus abyssi* genome sequence: insights into archaeal chromosome structure and evolution. Unpublished GenBank entry NC_000868 (▸http://www.ncbi.nlm.nih.gov), 2001.

Himmelreich, R., et al. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24,** 4420–4449 (1996).

Holt, R. A., et al. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298,** 129–149 (2002).

International Human Genome Sequence Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

Joyce, G. F. The antiquity of RNA-based evolution. *Nature* **418,** 214–221 (2002).

Kaiser, D. Building a multicellular organism. *Annu. Rev. Genet.* **35,** 103–123 (2001).

Kalman, S., et al. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat. Genet.* **21,** 385–389 (1999).

Kaneko, T., et al. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II.

Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3,** 109–136 (1996).

Katinka, M. D., et al. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi. Nature* **414,** 450–453 (2001).

Kawarabayasi, Y., et al. Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain 7. *DNA Res.* **8,** 123–140 (2001).

Kawarabayasi, Y., et al. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. *DNA Res.* **5,** 55–76 (1998).

Kawashima, T., et al. Determination of the complete genomic DNA sequence of *Thermoplasma volvanium* GSS1. *Proc. Jpn. Acad.* **75,** 213–218 (1999).

Klenk, H. P., et al. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus. Nature* **390,** 364–370 (1997).

Knapp, S., et al. Taxonomy needs evolution, not revolution. *Nature* **419,** 559 (2002).

Koonin, E. V. Genome sequences: Genome sequence of a model prokaryote. *Curr. Biol.* **7,** R656–659 (1997).

Koonin, E. V. Computational genomics. *Curr. Biol.* **11,** R155–158 (2001).

Korab-Laskowska, M., et al. The Organelle Genome Database Project (GOBASE). *Nucleic Acids Res.* **26,** 138–144 (1998).

Kumar, S. and Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392,** 917–920 (1998).

Kunst, F., et al. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis. Nature* **390,** 249–256 (1997).

Lander, E. S., and Waterman, M. S. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2,** 231–239 (1988).

Lang, B. F., Gray, M. W., and Burger, G. Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.* **33,** 351–397 (1999).

Lilly, J. W., and Havey, M. J. Small, repetitive DNAs contribute significantly to the expanded mitochondrial genome of cucumber. *Genetics* **159,** 317–328 (2001).

Mayr, E. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance.* Belknap Harvard, Cambridge, MA, 1982.

May, B. J., et al. Complete genomic sequence of *Pasteurella multocida,* Pm70. *Proc. Natl. Acad. Sci. USA* **98,** 3460–3465 (2001).

Nelson, K. E., et al. Evidence for lateral gene transfer between *Archaea* and bacteria from genome sequence of *Thermotoga maritima. Nature* **399,** 323–329 (1999).

Ng, W. V., et al. Genome sequence of *Halobacterium* species NRC-1. *Proc. Natl. Acad. Sci. USA* **97,** 12176–12181 (2000).

Nierman, W. C., et al. Complete genome sequence of *Caulobacter crescentus. Proc. Natl. Acad. Sci. USA* **98,** 4136–4141 (2001).

Ohyama, K., et al. Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. *Nature* **322,** 572–574 (1986).

Oliver, S. G., et al. The complete DNA sequence of yeast chromosome III. *Nature* **357,** 38–46 (1992).

Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276,** 734–740 (1997).

Parkhill, J., et al. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403,** 665–668 (2000a).

Parkhill, J., et al. Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404,** 502–506 (2000b).

Pennisi, E. Taxonomy. Linnaeus's last stand? *Science* **291,** 2304–2307 (2001).

Perna, N. T., et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409,** 529–533 (2001).

Read, T. D., et al. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* **28,** 1397–1406 (2000).

Ruepp, A., et al. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum. Nature* **407,** 508–513 (2000).

Sanger, F., et al. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265,** 687–895 (1977).

She, Q., et al. The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. *Proc. Natl. Acad. Sci. USA* **98,** 7835–7840 (2001).

Shinozaki, K. M., et al. The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. *EMBO J.* **5,** 2043–2049 (1986).

Simpson, A. J., et al. The genome sequence of the plant pathogen *Xylella fastidiosa. Nature* **406,** 151–157 (2000).

Skaletsky, H., et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423,** 825–837 (2003).

Smith, D. R., et al. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: Functional analysis and comparative genomics. *J. Bacteriol.* **179,** 7135–7155 (1997).

Stein, L. Genome annotation: From sequence to biology. *Nat. Rev. Genet.* **2,** 493–503 (2001).

Stephens, R. S., et al. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis. Science* **282,** 754–759 (1998).

Stover, C. K., et al. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406,** 959–964 (2000).

Takami, H., et al. Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with *Bacillus subtilis. Nucleic Acids Res.* **28,** 4317–4331 (2000).

Tettelin, H., Radune, D., Kasif, S., Khouri, H., and Salzberg, S. L. Optimized multiplex PCR: Efficiently closing a whole-genome shotgun sequencing project. *Genomics* **62,** 500–507 (1999).

Tettelin, H., et al. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287,** 1809–1815 (2000).

Tettelin, H., et al. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293,** 498–506 (2001).

Tomb, J. F., et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388,** 539–547 (1997).

Van de Peer, Y., De Rijk, P., Wuyts, J., Winkelmans, T., and De Wachter, R. The European small subunit ribosomal RNA database. *Nucleic Acids Res.* **28,** 175–176 (2000).

Venter, J. C., et al. The sequence of the human genome. *Science* **291,** 1304–1351 (2001).

Weber, J. L., and Myers, E. W. Human whole-genome shotgun sequencing. *Genome Res.* **7,** 401–409 (1997).

White, O., et al. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286,** 1571–1577 (1999).

Wilson, E. O. *The Diversity Of Life*. W. W. Norton, New York, 1992.

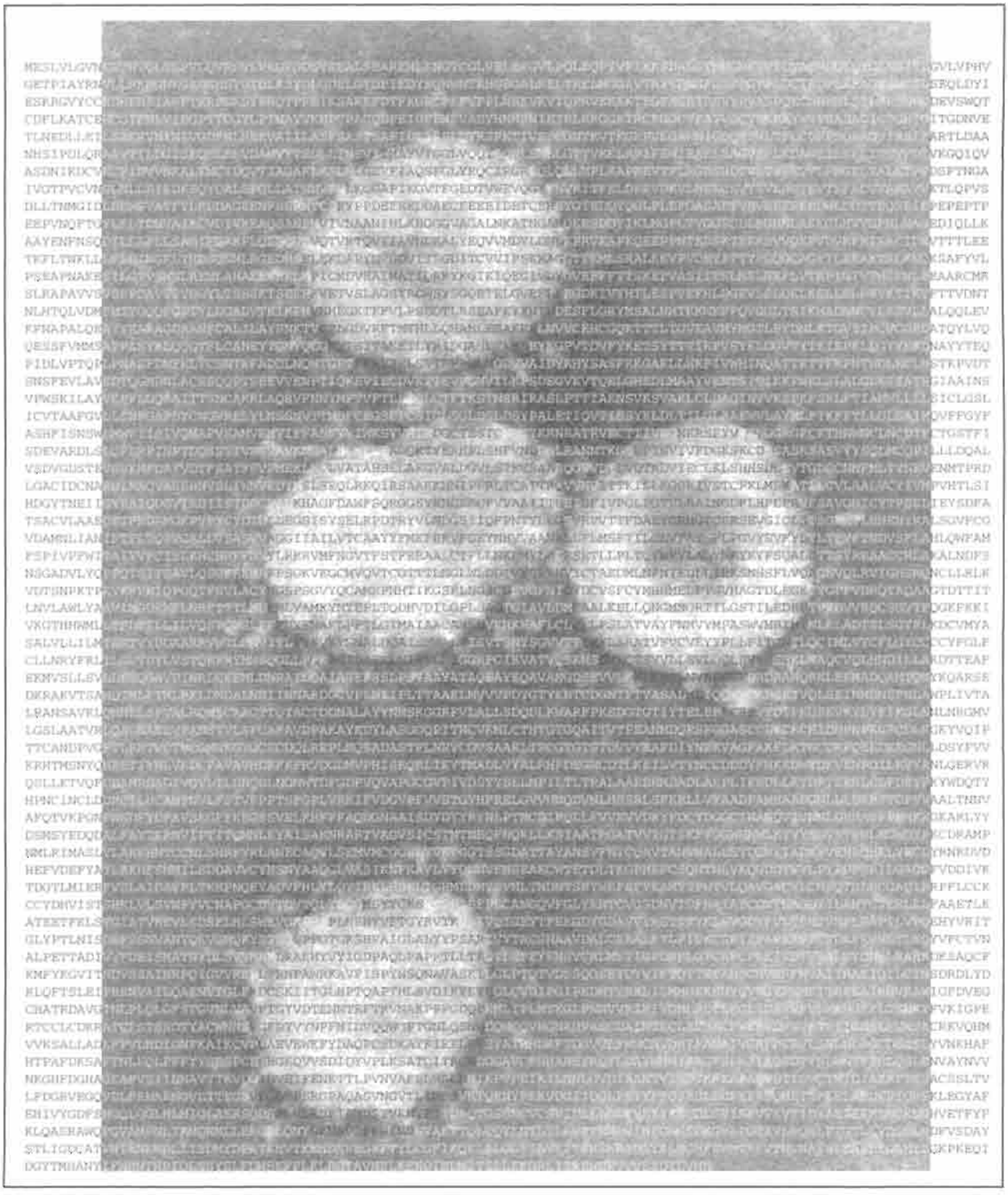
Woese, C. R. Default taxonomy: Ernst Mayr's view of the microbial world. *Proc. Natl. Acad. Sci. USA* **95,** 11043–11046 (1998).

Woese, C. R., Kandler, O., and Wheelis, M. L. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87,** 4576–4579 (1990).

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., and Koonin, E. V. Genome trees and the tree of life. *Trends Genet.* **18,** 472–479 (2002).

Wood, V., et al. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415,** 871–880 (2002).

Wuyts, J., Van de Peer, Y., Winkelmans, T., and De Wachter, R. The European database on small subunit ribosomal RNA. *Nucleic Acids Res.* **30,** 183–185 (2002).



Severe Acute Respiratory Syndrome (SARS) emerged in China in late 2002 and rapidly spread as a global epidemic. It is caused by a coronavirus, such as those shown in this image. The SARS genome (accession NC.004718) consists of 29,727 nucleotides. It was sequenced independently in a matter of weeks by groups in Canada and at the Centers for Disease Control and Prevention (CDC). One protein (accession NP.828849), shown here, is 7073 amino acids in length (predicted molecular mass about 790,000 daltons). A search of the Entrez protein database with the command "790000:999999[MolWt]" shows that only about one hundred proteins are currently known having a molecular weight of 790,000 daltons or more. The coronavirus image is from the CDC website (▸ http://phil.cdc.gov/Phil/detail.asp?id=3492).

13

Completed Genomes: Viruses

INTRODUCTION

In this chapter we will consider bioinformatic approaches to viruses. Viruses are small, infectious, obligate intracellular parasites. They depend on host cells for their ability to replicate. The virion (virus particle) consists of a nucleic acid genome surrounded by coat proteins (capsid) that may be enveloped in a lipid bilayer (derived from the host cell) studded with viral glycoproteins. Unlike other genomes, viral genomes can consist of either DNA or RNA. Furthermore, they can be single, double, or partially double stranded, and can be circular, linear, or segmented (having different genes on distinct nucleic acid segments).

Viruses lack the biochemical machinery that is necessary for independent existence. This is the fundamental distinction between viruses and free-living organisms. Thus, while they replicate and evolve, viruses exist on the borderline of the definition of life. The largest viruses (such as pox viruses) have genome sizes of several hundred kilobases that are nearly the same size as the smallest archaeal and bacterial genomes (such as *Nanoarchaeum equitans* and *Mycoplasma genitalium*) (Chapter 14). It is not a coincidence that those smallest prokaryotic genomes are from organisms that (like viruses) are small, infectious, obligate intracellular agents (see Chapter 14).

While there may be tens or hundreds of millions of species of bacteria and archaea, only a few thousand species of virus are known. This disparity probably reflects their specialized requirement for invading a host. Viruses infect all forms of

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

life, including bacteria, archaea, and eukaryotes from plants to humans to fungi.

Although viruses are relatively simple agents, they are more complex than two other pathogenic agents: viroids and prions. Viroids are small, circular RNA molecules of 200–400 nucleotides that cause diseases in plants (Flores, 2001). This minuscule genome does not encode any proteins, and the RNA itself has enzymatic activity. Prions are infectious protein molecules (Prusiner, 1998). Cruetzfeld–Jakob disease is the most common human prion disease (Johnson and Gibbs, 1998). It has a worldwide incidence of one in one million individuals and usually presents as dementia. Scrapie in sheep and bovine spongiform encephalopathy (BSE; "mad cow" disease) are the most common prion diseases in animals.

CLASSIFICATION OF VIRUSES

Before the sequencing era, morphology was an important criterion for the classification of viruses. Since 1959, electron microscopy has been employed to describe the structure of over 5100 bacteriophages (Ackermann, 2001) as well as additional viruses that invade plants and animals. Ninety-six percent of bacteriophages are tailed viruses, with the remainder having filamentous, icosahedral, or pleomorphic shapes. Many electron microscopic images of viruses are available at ICTVdb, the database of the International Committee on Taxonomy of Viruses (ICTV) (Büchen-Osmond, 1997). Several of these images are presented in Figure 13.1.

The ICTV website is at ▸http://www.ncbi.nlm.nih.gov/ICTVdb/index.htm. The ICTVdb was constructed by Cornelia Büchen-Osmond (Bioinformatics Group, Australian National University).

Another fundamental basis for classifying viruses is to define the type of nucleic acid genome that is packaged into the virion. Virions contain DNA or RNA; the nucleic acid may be single or double stranded, and translation may occur from the sense strand, the antisense strand, or both. Double-stranded viral genomes replicate by using the individual strands of the DNA or RNA duplex as a template to synthesize daughter strands. Single-stranded DNA or RNA viruses use their strand of nucleic acid as a template for a polymerase to copy a complementary strand. Replication may involve the stable or transient formation of double-stranded intermediates. Some viruses with single-stranded RNA genomes convert the RNA strand to DNA using reverse transcriptase (RNA-dependent DNA polymerase). In the case of HIV-1, the *pol* gene encodes reverse transcriptase.

The ICTV regularly meets to refine an accepted standard for virus classification. The ICTV recognizes the taxa of order, family, genus, and species. The ICTV database (sixth report) subdivides viruses into some 71 families, 9 subfamilies, and 164 genera [summarized by Mayo and Pringle (1998)]. An example of an online description of viruses on the ICTV website is provided in Figure 13.2.

Some of the major groups of viruses are shown in Figure 13.1 and Table 13.1. They range in genome size from very small viruses such as rubella (≈2 kb) to several viruses over 350 kb in size. A giant virus (called Mimivirus for *Mimicking microbe*) has been described, having a double-stranded circular genome of 800 kb (La Scola et al., 2003). Its mature particles are 400 nanometers in diameter. It is thus larger than many bacteria and archaea.

The Mimivirus GenBank accession number is AABV01000000.

An entirely different approach to classifying viruses is to identify those that cause human disease. Many viral diseases can be prevented by vaccination (Table 13.2). Others, such as smallpox, are of recent concern because of their potential use by bioterrorists (Cieslak et al., 2002). Smallpox, caused by the variola virus, was eradicated in 1977, and routine vaccination was discontinued in 1972 in the United States.

The National Institute of Allergy and Infectious Diseases (NIAID) at the National Institutes of Health offers information on viral and other diseases at ▸http://www.niaid.nih.gov/publications/.

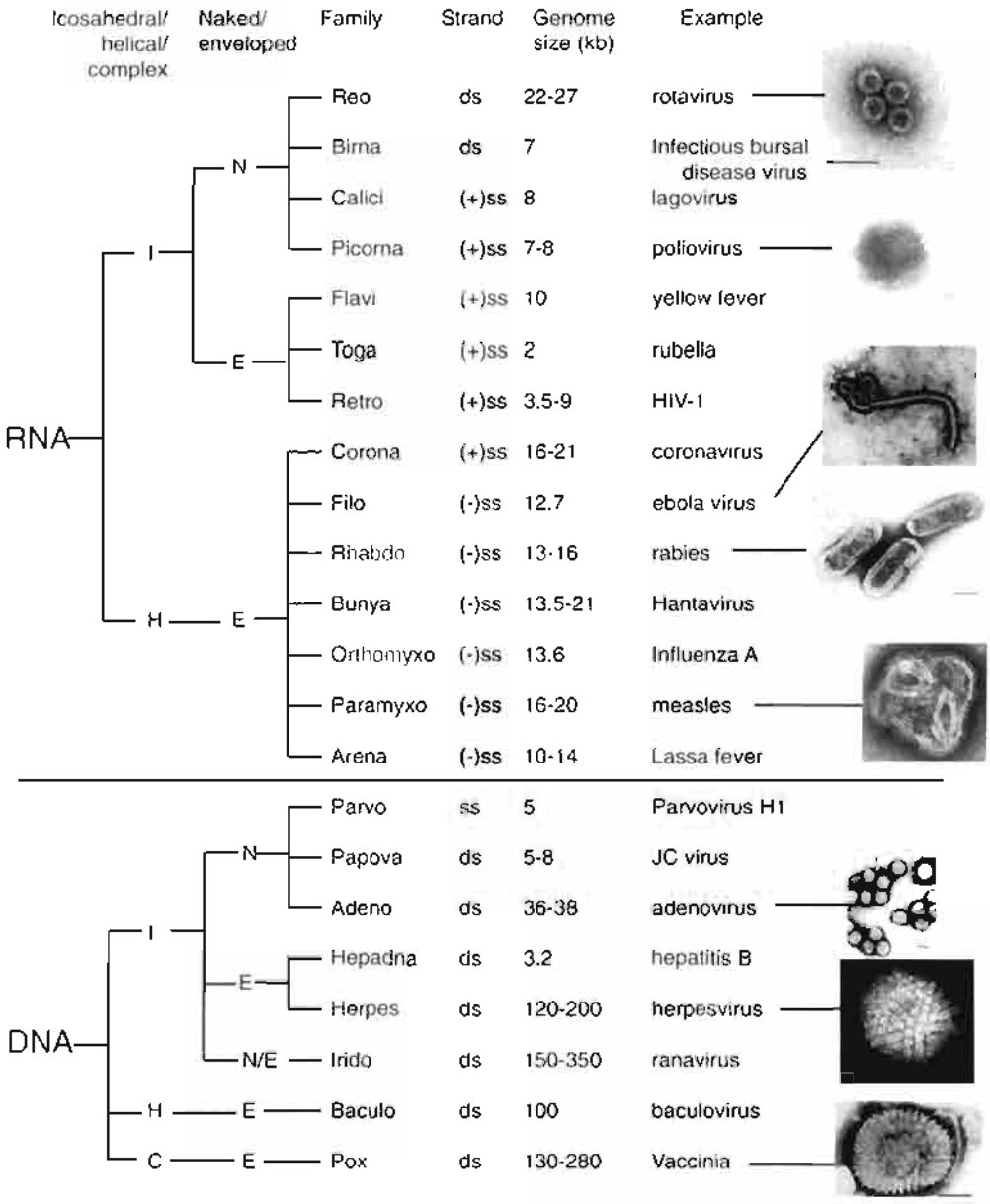


FIGURE 13.1. Classification of viruses. Adapted from ICTVdb and Flint et al. (2000, pp. 16–17). Electron micrographs are from the ICTV website.

BIOINFORMATICS APPROACHES TO PROBLEMS IN VIROLOGY

The tools of bioinformatics are well suited to address some of the outstanding problems in virology:

- Why does a virus such as HIV-1 infect one species selectively (human) while a closely related virus (simian immunodeficiency virus) infects monkeys but not humans? Analysis of the sequence of the viruses as well as the host cell receptors can address this question.
- Why do some viruses change their natural host? In 1997 a chicken influenza virus infected 18 humans, killing 6. Genes sequenced from an influenza A virus that infected a human were sequenced but did not appear to determine host specificity (Suarez et al., 1998).

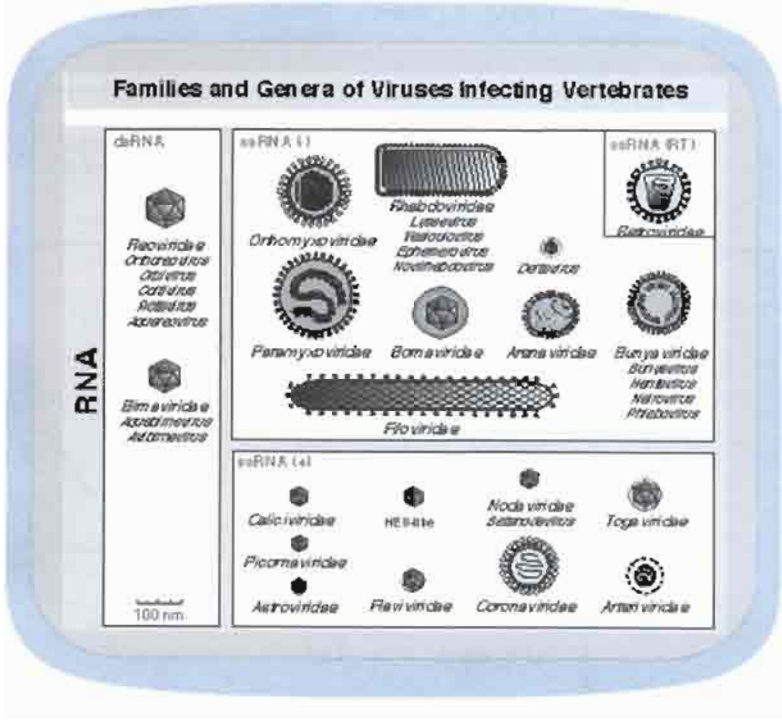


FIGURE 13.2. Example of a diagram illustrating virus morphology from the ICTV database (http://www.ncbi.nlm.nih.gov/ICTV/Diagrams/verrna.html).

- Why are some viral strains deadlier than others? An outbreak of influenza in 1918 killed an estimated 40 million to 100 million people worldwide in a single flu season. Researchers have collected 1918 virus samples to compare the virus sequence to modern influenza strains (Reid et al., 2002).
- What are the mechanisms of viral evasion of host immune systems? We will see below how some herpesviruses acquire viral homologs of human immune system molecules and thus interfere with human antiviral mechanisms.

TABLE 13-1 Classification of Viruses Based on Nucleic Acid Composition

Nucleic Acid	Strands	Family	Example	Accession	Base Pairs
RNA	Single	Picornaviridae	Human poliovirus 1	NC.002058	7,440
		Togaviridae	Rubella virus	NC.001545	9,755
		Flaviviridae	Yellow fever virus	NC.002031	10,862
		Coronaviridae	Coronavirus	NC.002645	27,317
		Rhabdoviridae	Rabies virus	NC.001542	11,932
		Paramyxoviridae	Measles virus	NC.001498	15,894
		Orthomyxoviridae	Influenza A virus (segment 1)	NC.002023	13,585
		Bunyaviridae	Hantavirus	–	–
		Arenaviridae	Lassa fever virus	J04324	3,402
		Retroviridae	HIV	NC.001802	9,181
	Double	Reoviridae	Rotavirus		
DNA	Single	Parvoviridae	Parvovirus H1	NC.001358	5,176
	Mixed	Hepadnaviridae	Hepatitis B	NC.001707	3,215
	Double	Papovaviridae	JC virus	NC.001699	5,130
		Adenoviridae	Human adenovirus, type 17	NC.002067	35,100
		Herpesviruses	Human herpesvirus 1	NC.001806	152,261
		Poxviridae	Vaccinia	NC.001559	191,737

Source: Adapted in part from Schaechter et al. (1999, p. 292). Used with permission.

TABLE 13-2 Vaccine-Preventable Viral Diseases

Disease	Virus	Comment
Hepatitis A	Hepatitis A virus	Causes liver disease
Hepatitis B	Hepatitis B virus	Causes liver disease
Influenza	Influenza type A or B	Causes 20,000 deaths per year (U.S.)
Measles	Measles virus	See below
Mumps	Rubulavirus	A disease of the lymph nodes
Poliomyelitis	Poliovirus (three serotypes)	Inflammation of the gray matter of the spinal cord; kills neurons
Rotavirus	Rotavirus	Most common cause of diarrhea in children; kills 600,000 children annually worldwide
Rubella	Genus Rubivirus	Also called German measles.
Smallpox	Variola virus	Eradicated in 1977
Varicella	Varicella-zoster virus	About 75% of all children contract varicella by age 15

Source: Adapted from ▸http://www.cdc.gov/nip/diseases/disease-chart-hcp.htm.

- Where did viruses originate? There are three main theories:
  1. The regressive theory suggests that viruses are derived from more complex intracellular parasites that eliminated many nonessential features.
  2. Viruses could be derived from normal cellular components that now replicate autonomously.
  3. Viruses could have coevolved with their host cells, possibly originating from self-replicating RNA molecules.

  Phylogenetic analyses could help resolve these theories.
- Which vaccines are most likely to be effective? There are two main approaches to developing vaccines for viruses that display a great amount of molecular sequence diversity. One approach is to select isolates of a particular subtype based on regional prevalence. A second approach is to deduce an ancestral sequence or a consensus sequence for use as an antigen in vaccine development (Gaschen et al., 2002). These approaches depend on molecular phylogeny.

Diversity and Evolution of Viruses

A premise of taxonomy is that it should represent phylogeny. In the case of viruses, their unique, elusive, and sometimes fragile nature makes it difficult to trace their evolution in as comprehensive a fashion as can be accomplished with archaea, bacteria, and eukaryotes. Like living organisms, viruses are subject to mutation (genetic variability) and selection. But viral genomes evolve far faster than cellular genomes and present special difficulties for evolutionary studies:

- Viruses tend not to survive in archeological or historical samples. There is considerable evidence for the existence of viruses over 10,000 years ago, based upon human skeletal remains, historical accounts, and other historical artifacts. However, ancient viral DNA or RNA has not been recovered.
- Viral polymerases of RNA genomes typically lack proofreading activity. This leads to a mutation rate that may be 1 million to 10 million times greater than

According to George Gaylord Simpson (1963, p. 7), "Species are groups of actually or potentially inbreeding populations, which are reproductively isolated from other such groups. An evolutionary species is a lineage (an ancestral-descendant sequence of populations) evolving separately from others and with its own unitary evolutionary role and tendencies."

that of DNA genomes (McClure, 2000). For viruses having DNA genomes, the mutation rates are typically 20- to 100-fold higher than that of the host cell.

- In addition to a high mutation rate, many viruses also have an extremely high rate of replication . A single cell can produce 10,000 poliovirus particles, and an HIV-infected individual can produce 10^9 virus particles per day. This can lead to the formation of quasi-species (a population of related but nonidentical viruses).
- Many viral genomes are segmented. This allows segments to be shuffled among progeny, producing a great diversity of viral subtypes (see HIV section below). This is seen in the influenza viruses that cause widespread sickness each year.
- Viruses are often subjected to intense selective pressures such as host immune responses or antiviral drug therapies. The rapid mutation rate of HIV-1 ensures that some versions of the virus are likely to contain mutations conferring resistance to retroviral drugs, and these HIV-1 molecules will be selected for.
- Viruses have evolved to invade diverse species across the entire tree of life: archaea, bacteria, and eukaryotes. Viruses that infect plants (e.g., tomato bushy stunt virus), animals (e.g., SV40, rhinovirus, and poliovirus), as well as bacteria (e.g., bacteriophage ϕX174) all share a "viral β-barrel" or "viral jelly roll" fold in the capsid protein structure (Hendrix, 1999). Unless a remarkable case of convergent evolution occurred, this suggests that these viruses are homologous. A group of reoviruses that infect both plants and animals has a characteristic double-stranded RNA genome packaged in an unusual capsid. They share these features in common with a family of bacteriophages (such as ϕ6), again suggesting homology between viruses that infect different branches of life (Hendrix, 1999). Notably sequence identity has not been detected in analysis of genes and/or proteins from these viral genomes, highlighting the rapid rate of viral genome evolution.

The great diversity of viral genomes precludes us from making comprehensive phylogenetic trees based upon molecular sequence data that span the entire universe of viruses. This reflects the complex molecular evolutionary events that form viral genomes (McClure, 2000).

For a variety of viral families, phylogenetic trees have been generated. These are indispensable in establishing the evolution, host specificity, virulence, and other biological properties of viral species. We will next examine phylogenetic reconstructions of the herpesviruses. Phylogenetic trees have been generated for HIV (see below) and for other viruses from measles to hepatitis.

From Phylogeny to Gene Expression: Bioinformatic Approaches to Herpesvirus

Herpesviruses are a diverse group of double-stranded DNA viruses that include herpes simplex, cytomegalovirus, and Epstein-Barr virus. As an example of viral phylogeny, McGeoch et al. (1995) analyzed eight well-conserved genes to deduce a phylogeny of the herpesviruses. Their phylogenetic reconstruction agrees with the major taxonomy of the herpesviruses in three groups: α-herpesviruses (formally called Alphaherpesvirinae), β-herpesviruses (Betaherpesvirinae), and γ-herpesviruses

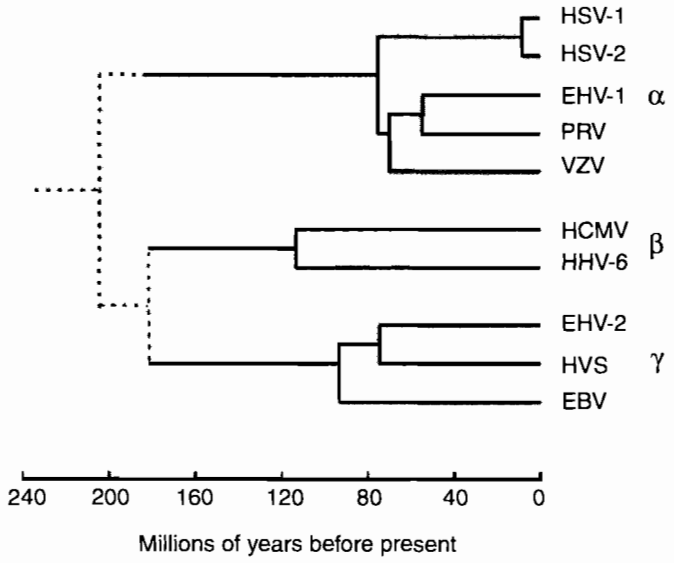


FIGURE 13.3. *Phylogeny of the herpesviruses. α-Herpesviruses include herpes simplex virus 1 and 2 (HSV-1 and HSV-2; the host is human for both), equine herpesvirus 1 (EHV-1), pseudorabies virus (PRV; the host is pig), and varicella-zoster virus (VZV; the host is human). β-Herpesviruses include human cytomegalovirus (HCMV) and human herpesvirus 6 (HHV-6). γ-Herpesviruses include equine herpesvirus 2 (EHV-2), herpesvirus saimiri (HVS; the host is the South American squirrel monkey), Epstein-Barr virus (EBV; the host is human), and human herpesvirus 8 (HHV-8; host human; not shown). The tree was generated from an alignment of conserved virion glycoproteins using the UPGMA method (Chapter 11). The dotted line indicates less confidence in dating the more ancient divergence times. From McGeoch et al. (1995). Used with permission.*

(Gammaherpesvirinae) (Fig. 13.3). This and similar analyses (Davison, 2002) provide great insight into the origin, diversity, and function of herpesviruses. Each herpesvirus is associated with a single host species (although some hosts, including humans, are infected by a variety of herpesviruses). This specificity suggests that herpesviruses have coevolved with their hosts over millions of years. Assuming a constant molecular clock, McGeoch et al. (1995) estimated that herpesviruses arose 180–220 MYA and that the three major sublineages arose before the time of mammalian radiation (and the end of the age of dinosaurs), 60–80 MYA.

Consider human herpesvirus 8 (HHV-8), a γ-herpesvirus. HHV-8 is also called Kaposi's sarcoma-associated herpesvirus, and it was initially identified by representational difference analysis in Kaposi's sarcoma lesions of AIDS patients (Chang et al., 1994). HHV-8 causes AIDS-associated Kaposi's sarcoma and other disorders such as primary effusion lymphoma and multicentric Castleman's disease. HHV-8 is closely related to rhesus rhadinovirus (RRV). The divergence of the HHV-8 and RRV may have coincided with speciation of humans and rhesus monkeys (Davison, 2002). The presence of both HHV-8 and an additional HHV-8-related virus in chimpanzees suggests that an additional virus may be identified that infects humans.

Kaposi's sarcoma is the most common tumor related to AIDS. It is a vascular malignancy that is typically first apparent in the skin.

What is the molecular basis for the cycle of latent and lytic infection by HHV-8? The genome is about 140,000 bp (NC_003409) and encodes over 80 proteins (Russo et al., 1996). We can explore the genome at the NCBI website using the Entrez genomes tool (Chapter 12). There are additional NCBI resources for the study of viruses. From a viral genomes home page (Fig. 13.4) you can link to double-stranded DNA viruses (such as the herpesviruses) or you can select the Clusters of Related Viral Proteins (CRP) tool (Fig. 13.5). From this site, you can browse the double-stranded DNA viruses (Fig. 13.6) and obtain a list of several dozen herpesvirus

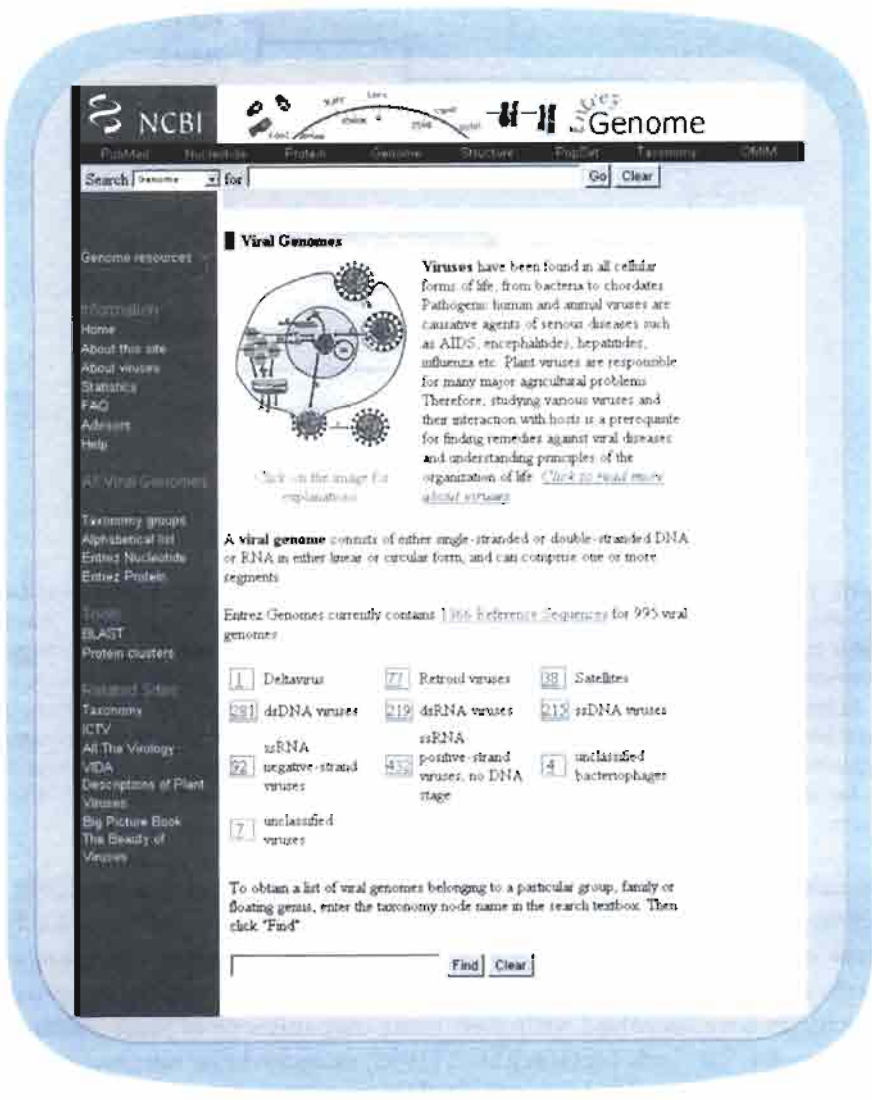


FIGURE 13.4. The viral genomes page at NCBI provides information and resources for the study of viruses.

genomes (Fig. 13.7). Select HHV-8 and you can view its open reading frames in a graphic form (Fig. 13.8) or a table (Fig. 13.9).

The HHV-8 proteins include virion structural and metabolic proteins. Interestingly, it also contains a variety of viral homologs of human host proteins such as complement-binding proteins, the apoptosis inhibitor Bcl-2, dihydrofolate reductase, interferon regulatory factors, an interleukin 8 (IL-8) receptor, a neural cell adhesion molecule–like adhesin, and a D-type cyclin.

How can viral genomes acquire a motif or an entire gene from a host organism? This can occur by a variety of mechanisms, including recombination, transposition, splicing, translocation, and inversion (McClure, 2000). Consider the IL-8 receptor, a eukaryotic gene that functions in cell growth and survival. This receptor is a member of the large family of G-protein-coupled receptors, including rhodopsin (that responds to light), the beta-adrenergic receptor (that binds adrenalin), and a variety of neurotransmitter receptors. Several viruses in addition to HHV-8 contain genes that encode a viral IL-8 receptor. A blastp search using this protein as a query (accession AAD46503) reveals that several viruses have proteins that are

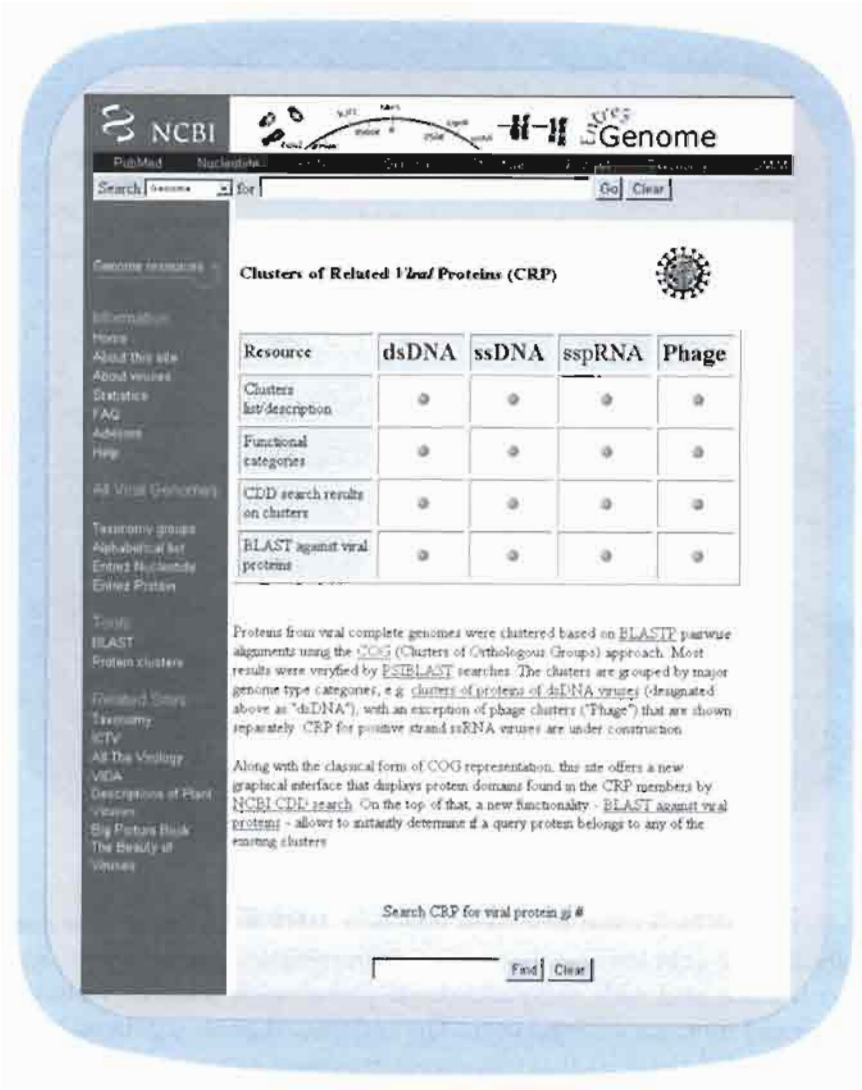


FIGURE 13.5. *The NCBI CRP website.*

distinct from but closely related to mammalian IL-8 receptors (Fig. 13.10). Presumably, when the virus infects a mammalian cell, this viral IL-8 protein is expressed and confers growth and survival that is advantageous to the virus (Wakeling et al., 2001).

Two complementary approaches have been taken to further study the function of viral genes (such as v-*IL-8* receptor) as well as mechanisms of HHV-8 infection. Paulose-Murphy et al. (2001) synthesized a microarray that represents 88 HHV-8 open reading frames and measured the transcriptional response of viral genes that are activated during the lytic replication cycle of HHV8 in human cells. They measured gene expression across a time series after inducing lytic infection and described clusters of genes that are coexpressed. Such genes may be functionally related. Clusters of genes coexpressed at early time points include several implicated in activation of the lytic viral cycle; another group of genes encode proteins that function in virion assembly (Fig. 13.11). The viral homologs of human proteins were expressed throughout the induced lytic cycle.

In an independent study, Poole et al. (2002) infected human dermal microvascular endothelial cells with HHV-8 and measured the transcriptional response of

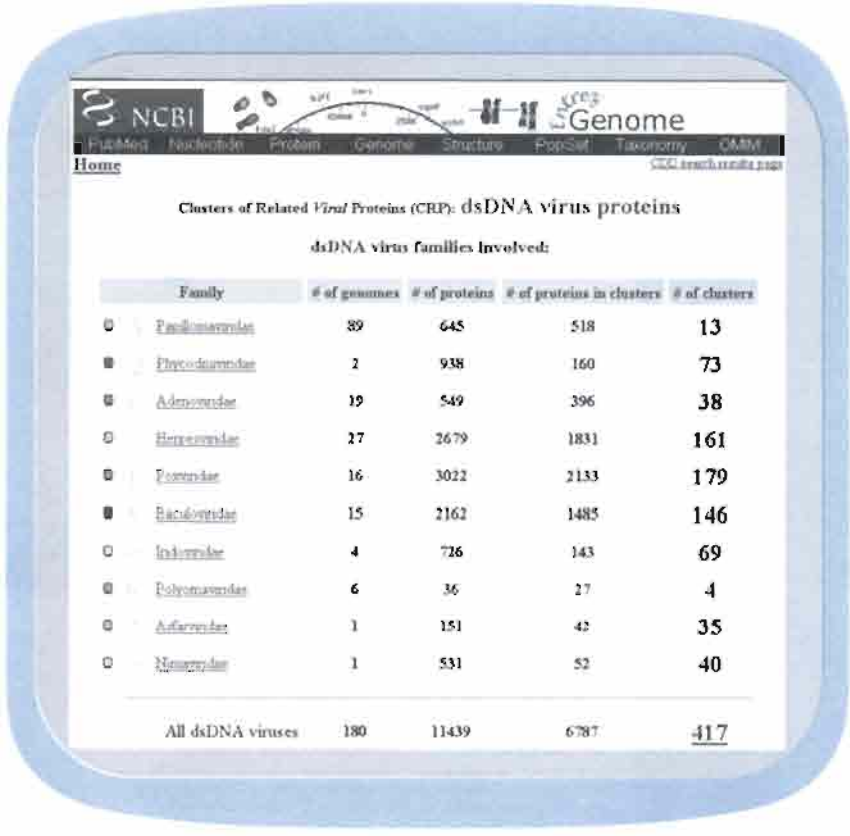


FIGURE 13.6. *From the NCBI CRP page, you can obtain a list of double-stranded DNA virus protein families. This includes the Herpesviridae.*

host cells to both latent and lytic virus infection. HHV-8 transforms the endothelial cells from a cobblestone shape to a characteristic spindle shape. Kaposi's sarcoma is associated with many additional pathological features, including angiogenesis and immune dysregulation. The endothelial genes regulated by HHV-8 infection included those such as interferon-responsive genes involved in immune function and genes encoding proteins with roles in cytoskeletal function, apoptosis, and angiogenesis. Such studies may be useful in defining the cellular response to viral infection.

Apoptosis is a type of programmed cell death in which the cell actively commits suicide. It serves as a mechanism by which a host cell can destroy infected cells, preventing a pathogen from spreading throughout the body. However, viruses have adapted to manipulate the cellular death pathway. Angiogenesis is the development of blood vessels. Infectious viruses (and cancerous tumors) require the presence of an adequate blood supply and sometimes promote angiogenesis.

Human Immunodeficiency Virus and Need for Bioinformatic Approaches

Human immunodeficiency virus is the cause of AIDS (reviewed in Meissner and Coffin, 1999). Until recently, HIV has been uniformly fatal. Most of the symptoms of AIDS are not caused directly by the virus but instead are a consequence of the ability of the virus to compromise the host immune system. Thus HIV infection leads to disease caused by opportunistic organisms.

At the end of the year 2002, 42 million people were infected with AIDS worldwide, and an additional 16 million people have died from AIDS. About 5 million new cases appeared in the year 2002. The prevalence of AIDS is increasing by about 3% per year.

HIV-1 and HIV-2 are retroviruses of the group lentivirus. The viruses probably originated in sub-Saharan Africa, where the diversity of viral strains is greatest and the infection rates are highest (Sharp et al., 2001). The primate lentiviruses occur

Information about AIDS is available at ▸http://www.niaid.nih.gov/factsheets/aidsstat.htm, an NIH website. Information on prevalence is from the Centers for Disease Control and Prevention at ▸http://63.126.3.84/2002/Abstract/13996.htm and UNAIDS and the World Health Organization at ▸http://www.unaids.org/.

FIGURE 13.7. The Herpesviridae page includes links to several dozen herpesvirus genomes.

in five major lineages, as shown by a phylogenetic tree based on full-length pol protein sequences (Fig. 13.12*a*; see arrows 1–5) (Hahn et al., 2000). These five lineages are:

1. Simian immunodeficiency virus (SIV) from the chimpanzee *Pan troglodytes* (SIVcpz), together with HIV-1
2. SIV from the sooty mangabeys *Cerecocebus atys* (SIVsm), together with HIV-2 and SIV from the macaques (genus *Macaca;* SIVmac)
3. SIV from African green monkeys (genus *Chlorocebus;* SIVagm)
4. SIV from Sykes' monkeys, *Cercopithecus albogularis* (SIVsyk)
5. SIV from l'Hoest monkeys, *Cercopithecus lhoesti* (SIVlhoest); SIV from suntailed monkeys (*Cercopithecus solatus;* SIVsun); and SIV from a mandrill (*Mandrillus sphinx;* SIVmnd)

A prominent feature of phylogenetic analyses such as those in Figure 13.12*a* is that viruses appear to have evolved in a host-dependent manner (Hahn et al., 2000). Viruses infecting any particular nonhuman primate species are more closely related to one another than they are related to viruses from other species. For HIV-2,

Prevalence of a disease (or infection) is the proportion of individuals in a population who have a disease at a particular time. Prevalence does not describe when individuals contracted a disease. Incidence is the frequency of new cases of a disease that occur over a particular time. For example, the incidence of a disease might be described as 10 new cases per 1000 people in the general population in a given year.

FIGURE 13.8. The HHV-8 genome at NCBI.

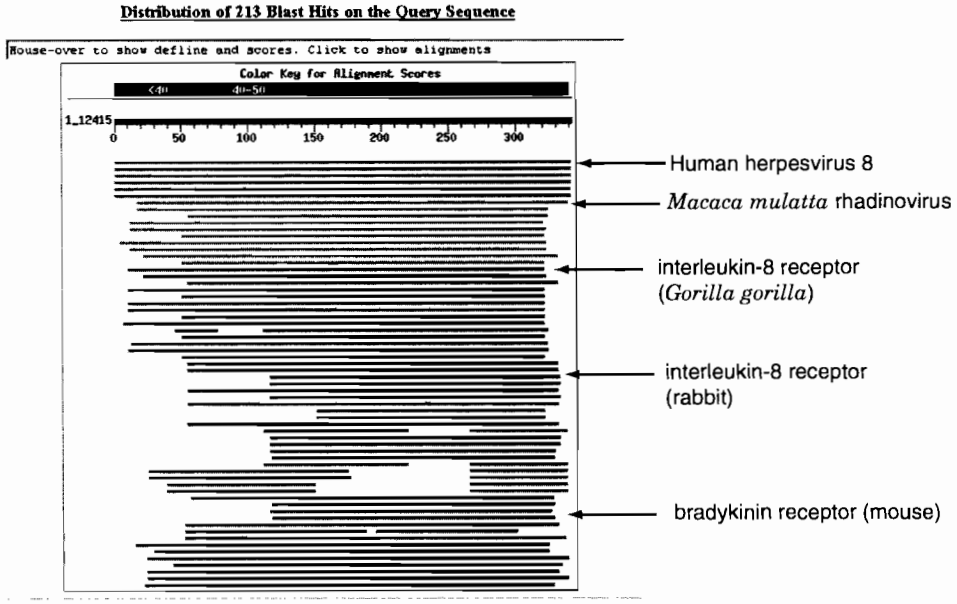
Human herpesvirus 8, genome

Send the report below in Table format

◆ - GenBank record including protein ◇ - DNA region in flatfile format ◆ - DNA and protein in FASTA format

	Location	Strand	Length	PID	Gene	Synonym	Product
◆ ◇ ◆	105..974	+	290	10845966			ORF K1
◆ ◇ ◆	1142..2794	+	551	10845967			ORF 4; The HVS ORF 4 homolog has alternativel
◆ ◇ ◆	3210..4611	+	1134	10845968			ORF 6; ss DNA binding protein ssDBP homolog;
◆ ◇ ◆	6628..8715	+	696	10845969			ORF 7; transport protein homolog; EBV BALF3 h
◆ ◇ ◆	8699..11234	+	846	10845970			ORF 8; glycoprotein B gB homolog; EBV BALF4 h
◆ ◇ ◆	11365..14401	+	1013	10845971			ORF 9; DNA polymerase homolog; EBV BALF5 homo
◆ ◇ ◆	14519..15775	+	419	10845972			ORF 10
◆ ◇ ◆	15790..17013	+	408	10845973			ORF 11; EBV Raji LF2 homolog
◆ ◇ ◆	17261..17875	-	205	10845974			ORF K2; functional interleukin-6 vIL-6 homolo
◆ ◇ ◆	17921..18553	-	211	10845975			ORF 2; dihydrofolate reductase DHFR homolog
◆ ◇ ◆	18608..19609	-	334	10845976			ORF K3; BHV4-IE1 homolog
◆ ◇ ◆	20091..21104	-	338	10845977			ORF 70; thymidylate synthase TS homolog
◆ ◇ ◆	21548..21832	-	95	10845978			ORF K4; macrophage inflammatory protein vMIP-
◆ ◇ ◆	25713..26483	-	257	10845979			ORF K5; BHV4-IE1 homolog
◆ ◇ ◆	27137..27424	-	96	10845980			ORF K6; functional macrophage inflammatory pr
◆ ◇ ◆	28622..29002	+	127	10845981			ORF K7
◆ ◇ ◆	30145..30672	+	176	10845982			ORF 16; functional anti-apoptotic factor vBCL
◆ ◇ ◆	30821..32482	-	554	10845983			ORF 17; minor capsid protein homolog; EBV BVR
◆ ◇ ◆	32424..33197	+	258	10845984			ORF 18
◆ ◇ ◆	33194..34843	-	550	10845985			ORF 19; tegument protein homolog; EBV BVRF1 h
◆ ◇ ◆	34611..35573	-	321	10845986			ORF 20; EBV BXRF1 homolog
◆ ◇ ◆	35383..37125	+	581	10845987			ORF 21; thymidine kinase TK homolog; EBV BXLF
◆ ◇ ◆	37113..39305	+	731	10845988			ORF 22; glycoprotein H gH homolog; EBV BXLF2
◆ ◇ ◆	39302..40516	-	405	10845989			ORF 23; EBV BTRF1 homolog
◆ ◇ ◆	40520..42778	-	753	10845990			ORF 24; EBV BcRF1 homolog
◆ ◇ ◆	42777..46907	+	1377	10845991			ORF 25; major capsid protein MCP homolog; EBV
◆ ◇ ◆	46933..47850	+	306	10845992			ORF 26; minor capsid protein homolog; EBV BDLF
◆ ◇ ◆	47873..48745	+	291	10845993			ORF 27; EBV BDLF2 homolog
◆ ◇ ◆	48991..49299	+	103	10845994			ORF 28; EBV BDLF3 homolog
◆ ◇ ◆	49362..50417	-	352	10845995			ORF 29b; packaging protein homolog; EBV BDRF1
◆ ◇ ◆	50623..50856	+	78	10845996			ORF 30; EBV BDLF3.5 homolog
◆ ◇ ◆	50769..51437	+	225	10845997			ORF 31; EBV BDLF4 homolog
◆ ◇ ◆	51404..52769	+	455	10845998			ORF 32; EBV BGLF1 homolog
◆ ◇ ◆	52761..53699	+	313	10845999			ORF 33; EBV BGLF2 homolog
◆ ◇ ◆	53738..54676	-	313	10846000			ORF 29a; packaging protein homolog; EBV BGRF1
◆ ◇ ◆	54675..55658	+	328	10846001			ORF 34; EBV BGLF3 homolog
◆ ◇ ◆	55639..56091	+	151	10846002			ORF 35; EBV BGLF3.5 homolog
◆ ◇ ◆	55976..57310	+	445	10846003			ORF 36; kinase homolog; EBV BGLF4 homolog
◆ ◇ ◆	57273..58733	+	487	10846004			ORF 37; alkaline exonuclease homolog; EBV BGL
◆ ◇ ◆	58680..58871	+	62	10846005			ORF 38; EBV BDLF1 homolog

FIGURE 13.9. The HHV-8 genome page includes a link to this table of 82 open reading frames.



Sequences producing significant alignments:	Score (bits)	E Value
gi|3183195|sp|Q98146|VG74_KSHV G-PROTEIN COUPLED RECEPTOR H...	538	e-152
gi|5929959|gb|AAD56655.1|AF179928_1 (AF179928) IL-8 recepto...	537	e-152
gi|13386561|gb|AAC28486.1| (AF079845) G-protein coupled rece...	536	e-152
gi|1930014|gb|AAB51506.1| (U82242) chemokine receptor homol...	535	e-151
gi|4154096|gb|AAD04749.1| (U40394) G-protein coupled recept...	535	e-151
gi|5929958|gb|AAD56654.1|AF179927_1 (AF179927) IL-8 recepto...	535	e-151
gi|4494986|gb|AAD21408.1| (AF083501) G protein coupled rece...	186	2e-46
gi|9626030|ref|NP_040276.1| G protein-coupled receptor hom...	102	4e-21
gi|9631265|ref|NP_048046.1| orf 74; IL-8 receptor family [A...	98	1e-19
gi|2494967|sp|Q28422|IL8B_GORGO HIGH AFFINITY INTERLEUKIN-8...	86	5e-16
gi|2494969|sp|Q28807|IL8B_PANTR HIGH AFFINITY INTERLEUKIN-8...	84	2e-15
gi|124357|sp|P21109|IL8A_RABIT HIGH AFFINITY INTERLEUKIN-8 ...	83	3e-15
gi|4504683|ref|NP_001548.1| interleukin 8 receptor, beta [H...	82	5e-15

FIGURE 13.10. *A viral protein is a G-protein coupled receptor that is homologous to a superfamily of mammalian G-protein coupled receptors, including a high-affinity interleukin 8 (IL-8) receptor. This protein is present in several viruses, including Kaposi's sarcoma-associated herpesvirus (also called HSV-8) and a murine γ-herpesvirus. The gene encoding this receptor was presumably of mammalian origin and integrated into the genomes of several viruses. Upon viral infection, this receptor may promote growth and survival of infected cells.*

transmission from the sooty mangabeys was indicated by five lines of evidence (Hahn et al., 2000):

1. Similarities in the genome structures of HIV-2 and SIVsm
2. Phylogenetic relatedness of HIV-2 and SIVsm (see Fig. 13.12, arrow 4)
3. Prevalence of SIVsm in the natural host
4. Geographic coincidence of those affected and the natural host
5. Plausible routes of transmission, such as exposure of humans to chimpanzee blood in markets

Similar arguments have been applied to HIV-1, which probably appeared in Africa in 1930–1940 as a cross-species contamination by SIVcpz. HIV-1 occurs in three major subtypes, called M, N, and O. This is consistent with the occurrence of three separate SIVcpz transmissions to humans: M is the main group of HIV-1 viruses; O is an outlier group; and N is also distinct from M and O. The three main HIV-1 subtypes are apparent in a phylogenetic tree generated from full-length Env protein sequences (Fig. 13.12*b*, arrows 6-8) (Hahn et al., 2000).

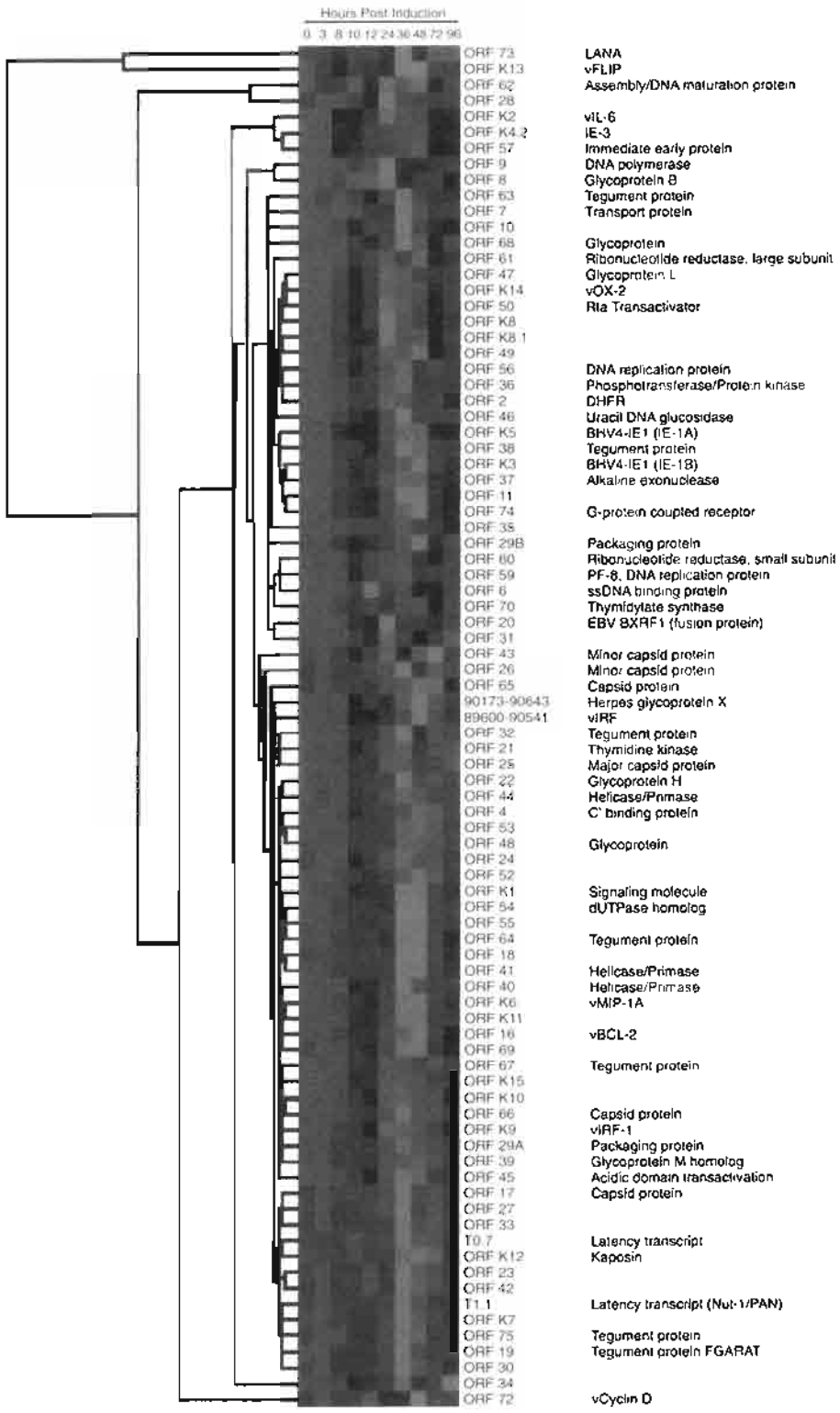


FIGURE 13.11. Two-way hierarchical clustering of microarray data using an HHV-8 array to measure HHV-8 gene expression in infected human cells. The temporal expression ratios of genes were compared pairwise and grouped according to their similarity. The columns indicate separate time points (hours postinduction of the viral lytic cycle). The row displays the expression profile of each single open reading frame. The normalized expression ratios across all time points are color coded to denote the level of up or down regulation. The dendrogram at left groups genes based on similar gene expression patterns across time. This approach is useful to discern the function of viral genes during infection. From Paulose-Murphy et al. (2001). Used with permission.

FIGURE 13.12. *Evolutionary relationships of primate lentiviruses.* (a) *Full-length Pol protein sequences were aligned and a tree was created using the maximum-likelihood method. There are five major lineages (arrows 1–5). The scale bar indicates 0.1 amino acid replacements per site after correction for multiple hits.* (b) *The HIV-1/SIVcpz lineage is displayed based on a maximum-likelihood tree using Env protein sequences. Note that the three major HIV-1 groups (M, N, O; arrows 6–8) are distinguished. The scale bar is the same as in* (a). *From Hahn et al. (2000). Used with permission.*

We saw the NCBI entry for HIV-1 in Figure 2.12; the genome is 9181 bases and encodes nine proteins. While the HIV-1 genome is small and there are few gene products, GenBank will soon have over 100,000 nucleotide sequence records and an equal number of protein records. The reason for this enormous quantity of data is that HIV-1 mutates extremely rapidly, producing many subtypes of the M, N, and O variants. Thus, researchers sequence HIV variants very often. A major challenge for virologists is to learn how to manipulate such large amounts of data and how to

As of June 2003, GenBank contains about 90,000 nucleotide records for HIV-1. To see this, go to the Taxonomy browser page and enter HIV-1. If you limit the output in a search of the Entrez nucleotide database to RefSeq entries, there is only one entry: the complete HIV-1 genome (NC.001802).

The Entrez Genome section (▸http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome) includes a listing of hundreds of viruses. As of June 2003, there are over 1300 completed virus genome sequences available at this site. Additionally there are over 100 phage sequences and dozens of viroids (infectious agents with RNA genomes that cause diseases in plants).

Under the virus category of "Entrez genomes" a link is provided to this virus (listed alphabetically) and to the HIV-1 accession number (NC.001802). By clicking on the name of the virus, one is linked to the NCBI taxonomy browser, which includes information on the lineage of HIV-1 (Viruses; Retroid viruses; Retroviridae; Lentivirus; Primate lentivirus group) as well as links to dozens of HIV-1 variants. From the Entrez Genome page, by clicking on the accession number NC_001802, one links to the Entrez Nucleotide (GenBank) entry for HIV-1. As indicated in the entry, this is a 9181-base single-stranded RNA molecule. On the left sidebar, by clicking "Coding Regions," one links to a table listing the coding regions in the virus. This is a convenient way to obtain the DNA (or amino acid) sequence corresponding to a specific gene (or protein) of interest.

From the NCBI home page (▸http://www.ncbi.nlm.nih.gov/), select "Retrovirus Resources" (▸http://www.ncbi.nlm.nih.gov/retroviruses/).

use those data to find meaningful approaches to treating or curing AIDS. We will next describe two bioinformatics resources for the study of HIV molecular sequence data: NCBI and LANL.

Bioinformatic Approaches to HIV-1

Using NCBI Resources

The NCBI website offers several ways to study retroviruses, including HIV. You can access information on HIV-1 via the Entrez Genome site at NCBI, as we have described for HHV-8 above.

NCBI also offers a specific resource for the study of retroviruses (Fig. 13.13). This site includes the following:

- A genotyping tool based upon BLAST searching
- A multiple sequence alignment tool specific for retroviral sequences
- A reference set of 50 retroviral genomes
- Specific pages with tools to study HIV-1, HIV-2, SIV, human T-cell lymphotropic virus type 1 (HTLV), and STLV
- A listing of the previous week's publications on retroviruses
- A listing of the previous week's GenBank releases (many hundreds of new HIV-1 sequences are deposited weekly)
- Links to external retroviral website resources.

FIGURE 13.13. Retroviruses resource from NCBI (▸http://www.ncbi.nlm.nih.gov/retroviruses/).

Using LANL Database

The Los Alamos National Laboratory (LANL) operates a group of four HIV databases with an associated website. The HIV Sequence Database is an important, comprehensive repository of HIV sequence data. It allows searches for sequences by common names, accession number, PubMed identifier, country in which each case was sampled, and likely country in which infection occurred. Sequences may be retrieved as part of a multiple sequence alignment or unaligned, and groups of sequences derived from an individual patient may be retrieved. The site includes a variety of specialized tools, including:

The LANL HIV databases are available at ▸http://hiv-web.lanl.gov/. This site offers four databases: sequence, resistance, immunology, and vaccine trials. In the HIV Sequence Database, you can find the geography tool at ▸http://hiv-web.lanl.gov/content/hiv-db/geography/geography.comp.

- An HIV BLAST server
- SNAP (Synonymous/Non-synonymous Analysis Program), a program that calculates synonymous and nonsynonymous substitution rates
- Recombinant Identification Program (RIP), a program that identifies mosaic viral sequences that may have arisen through recombination
- A multiple alignment program called MPAlign (Gaschen et al., 2001) that uses HMMER software (Chapter 10)
- PCoord (Principal Coordinate Analysis), a program that performs a procedure similar to principal components analysis (Chapter 7) on sequence data based on distance scores
- A geography tool that shows both total HIV infection levels (either worldwide or by continent) as well as the subtype distribution of HIV (Figs. 13.14*a*, *b*)

The LANL website includes other databases that provide important tools for the bioinformatic analysis of HIV-1 and related viruses. The HIV Drug Resistance Database allows you to browse HIV-1 genes to identify specific drugs that are affected by amino acid substitutions in that gene product (Fig. 13.15). This information is also displayed graphically with the mapping tool in the HIV Drug Resistance Database. This clickable plot shows the complete amino acid sequence of an HIV protein (e.g., the protease; Fig. 13.16*a*, arrow 1). Each row contains a different drug, and amino acid substitutions associated with the resistance of HIV to each of the drugs are also displayed. Clicking on a substituted amino acid (Fig. 13.16, arrow 2) leads to a report describing the mutation in detail (Fig. 13.16*b*).

The LANL Protease Mutations-by-Drug Map is available at ▸http://resdb.lanl.gov/Resist_DB/protease_mutation_map.htm.

The Frederick Cancer Research and Development Center of the National Cancer Institute has developed an HIV protease database (▸http://mcl1.ncifcrf.gov/hivdb/index.html) consisting of several hundred HIV-1, HIV-2, and SIV protease structures.

Bioinformatic Analysis of Viral Genome: Measles Virus

Measles virus is one of the deadliest viruses in human history. Today, it is the leading cause of death in children in many countries, killing over 1 million infants each year (Johnson et al., 2000). Vaccines have helped to reduce the mortality and morbidity rates, but the presence of an immature immune system and maternal antibodies prevent successful immunization in newborns before nine months of age. The virus spreads by respiratory droplets, infecting epithelial cells in the respiratory tract.

Before the measles vaccination was introduced in the United States, there were 450,000 cases annually (and about 450 deaths). See ▸http://www.cdc.gov/nip/diseases/measles/q&a.htm.

The measles virus is a Morbillivirus of the Paramyxoviridae family, which includes mumps and respiratory syncytial virus. You can access the genome through the NCBI Entrez genomes resource (accession NC_001498). Measles virus consists of a nonsegmented, negative sense RNA genome protected by nucleocapsids and an envelope. The genome has 15,894 bp and encodes seven proteins. These sequences can be accessed by clicking on the "coding regions" option on the left sidebar of the Entrez record (Fig. 13.17*a*). Six genes are designated N (nucleocapsid),

Another member of the Paramyxoviridae family is the cause of rinderpest, an ancient plague of cattle (Barrett and Rossiter, 1999). These viruses have had a devastating impact on both humans and ruminants.

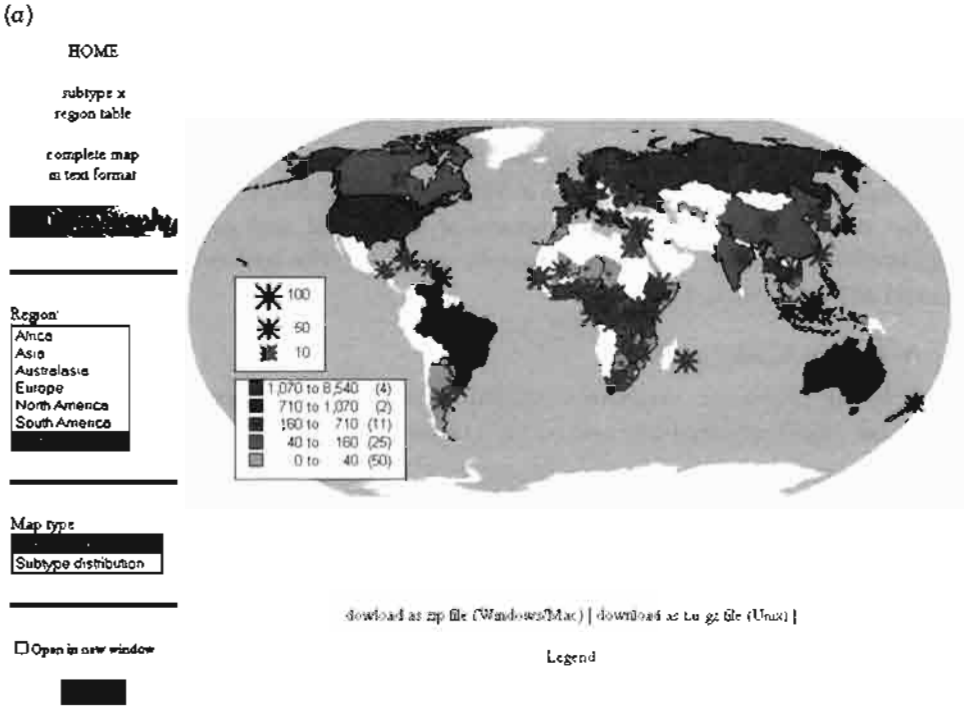
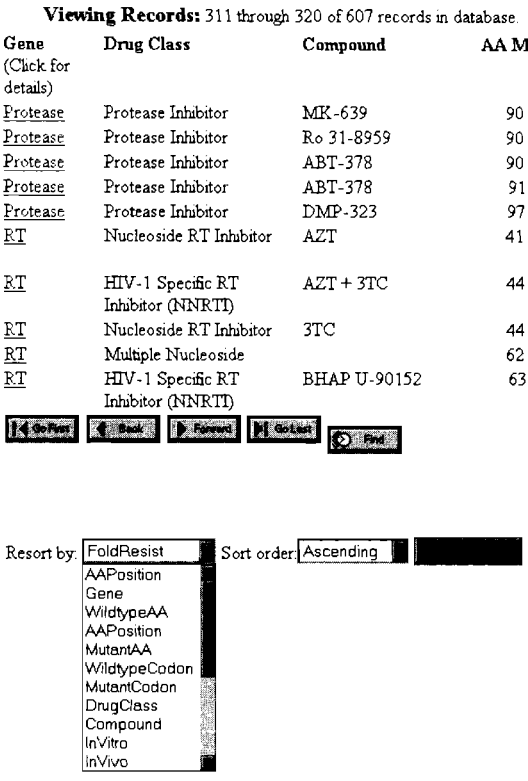


FIGURE 13.14. The geography tool at LANL allows you to view HIV infection subtypes (a) *globally or* (b) *by continent (Africa is shown). In* (a), *the total and dominant subtypes are indicated. The squares indicate the number of HIV sequences per country in the LANL database, and the stars indicate the percentage of nondominant sequences per country (a large star indicates a greater diversity of subtypes). The subtype distribution is displayed in* (b) *using pie charts. This geography tool is available at* ▸*http://hiv-web.lanl.gov/content/hiv-db/geography/map.right.html.*

P (phosphoprotein), M (matrix), F (fusion), H (hemagglutinin), and L (large polymerase) (Fig. 13.17*b*). The P gene is predicted to encode another protein (nonstructural C protein) using an alternative start site on a different reading frame. It is easy to visualize this by clicking on the Entrez nucleotide record, then choosing the "Graphics" display option (Fig. 13.18). This shows where the measles virus genome encodes the nonstructural C protein.

The functions of the six measles virus proteins have been assigned: N binds to genomic RNA and surrounds it, P and L form a complex involved in RNA synthesis,



Viewing Records: 311 through 320 of 607 records in database.

Gene (Click for details)	Drug Class	Compound	AA Mutation	Codon Mutation	Cite
Protease	Protease Inhibitor	MK-639	90 L M	TTG -> ATG	Condra96
Protease	Protease Inhibitor	Ro 31-8959	90 L M	TTG -> ATG	Jacobsen94
Protease	Protease Inhibitor	ABT-378	90 L M	TTG -> ATG	Kempf01
Protease	Protease Inhibitor	ABT-378	91 T S	ACT -> TCT	Carrillo98
Protease	Protease Inhibitor	DMP-323	97 L V	TTA -> GTA	King95
RT	Nucleoside RT Inhibitor	AZT	41 M L	ATG -> TTG/CTG	Larder89 Larder91 Kellam92
RT	HIV-1 Specific RT Inhibitor (NNRTI)	AZT + 3TC	44 E D	GAA -> GAC	Hertogs00
RT	Nucleoside RT Inhibitor	3TC	44 E A	GAA -> GCA	Montes02
RT	Multiple Nucleoside		62 A V	GCC -> GTC	Iversen96 Shirasaka95
RT	HIV-1 Specific RT Inhibitor (NNRTI)	BHAP U-90152	63 I M	ATA -> ATG	Pelemans01

Resort by: FoldResist Sort order: Ascending

AAPosition
Gene
WildtypeAA
AAPosition
MutantAA
WildtypeCodon
MutantCodon
DrugClass
Compound
InVitro
InVivo

FIGURE 13.15. *The HIV Drug Resistance Database compiles amino acid substitutions in HIV genes that confer resistance to anti-HIV drugs. The browse tool (available via ▸ http://resdb.lanl.gov/Resist_DB/default.htm) allows you to search any HIV-1 gene for substitutions. The data can be sorted by many criteria as shown in the pull-down menu.*

M links the ribonucleoprotein to the envelope glycoproteins H and F which are inserted in the virus membrane on the surface of the virion, H binds the cell surface receptor through which the virus enters its host, and F is a fusion protein that promotes insertion of the virus into the host cell membrane. The functions of each of these proteins can be assessed by performing BLAST searches. For the nonstructural C protein, a blastp nonredundant (nr) search reveals homology to proteins encoded by the genomes of rinderpest virus, canine and phocine distemper virus, and dolphin morbillivirus. A blastp nr search with the viral hemagglutinin reveals membership in a Pfam family (pfam00423, Hemagglutinin-neuraminidase), and there are several hundred matches to measles virus hemagglutinin. Repeat the search with the Entrez limit "hemagglutinin NOT measles virus[Organism]" and the results are reduced to several dozen hemagglutinins from the homologous morbilliviruses other than measles. A PSI-BLAST search identifies hundreds of additional hemagglutinins from viruses such as human parainfluenza, mumps, and a turkey rhinotracheitis virus.

Additional Bioinformatics Resources

Several specialized databases have been established for the study of viruses (see Table 13.3 under Web Resources). The EMBL Virus Structure Resource provides images of virus structures. Project VirgO offers software tools, including the Viral

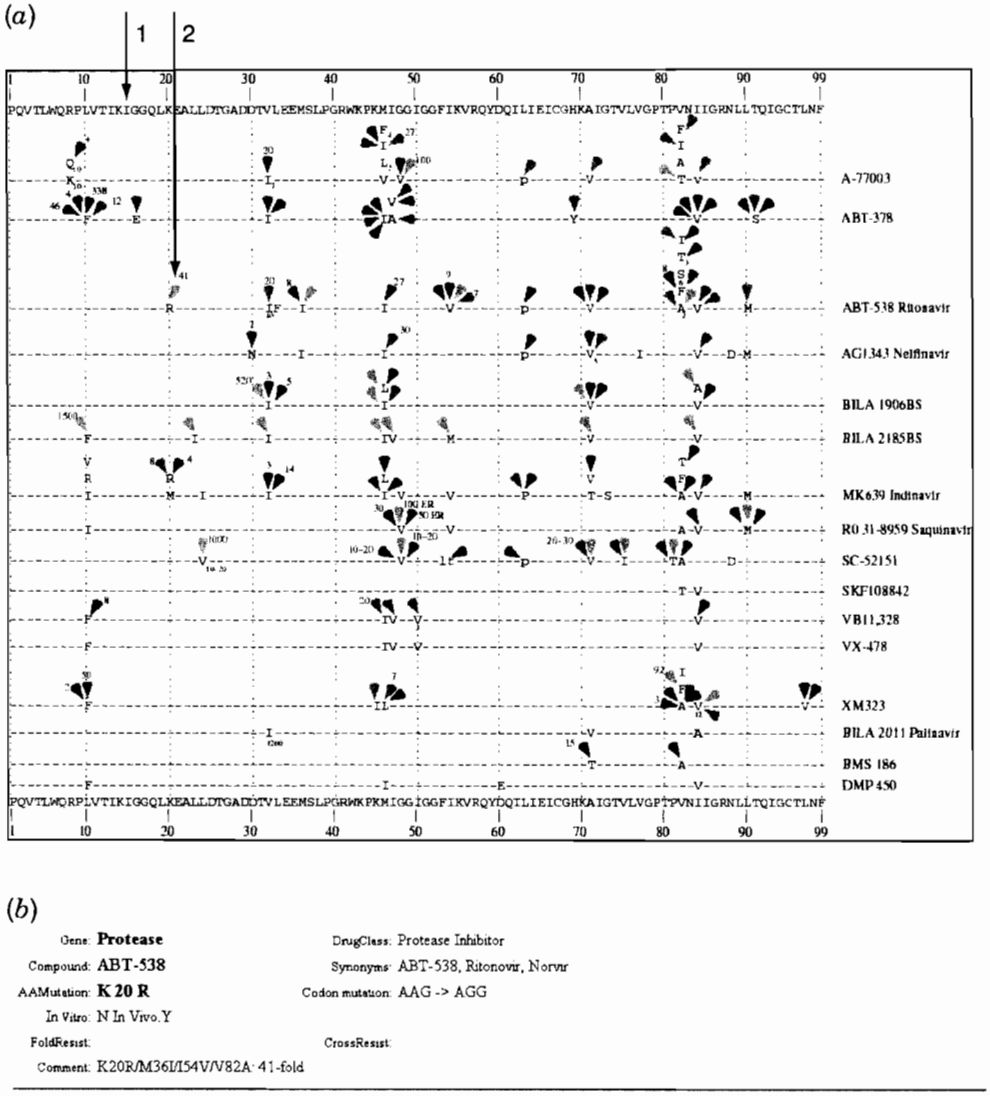


FIGURE 13.16. (a) *The LANL website offers a map of HIV-1 protease mutations versus drugs. Each row represents a drug (labeled at right). The wild-type (strain HXB2) HIV-1 protease sequence is listed at top and bottom (arrow 1). Dashes indicate wild-type amino acid positions, while mutations that confer resistance to the drug are indicated. An example of a K-to-R (lysine-to-arginine) mutation is indicated (arrow 2). The small number (41) indicates the "fold resistance" of that particular mutation. Mutations that have a colored shape pointing to them are also part of a synergistic combination of mutations.* (b) *By clicking on the position of a mutation (arrow 2), the map links to a detailed report of the effects of that mutation.*

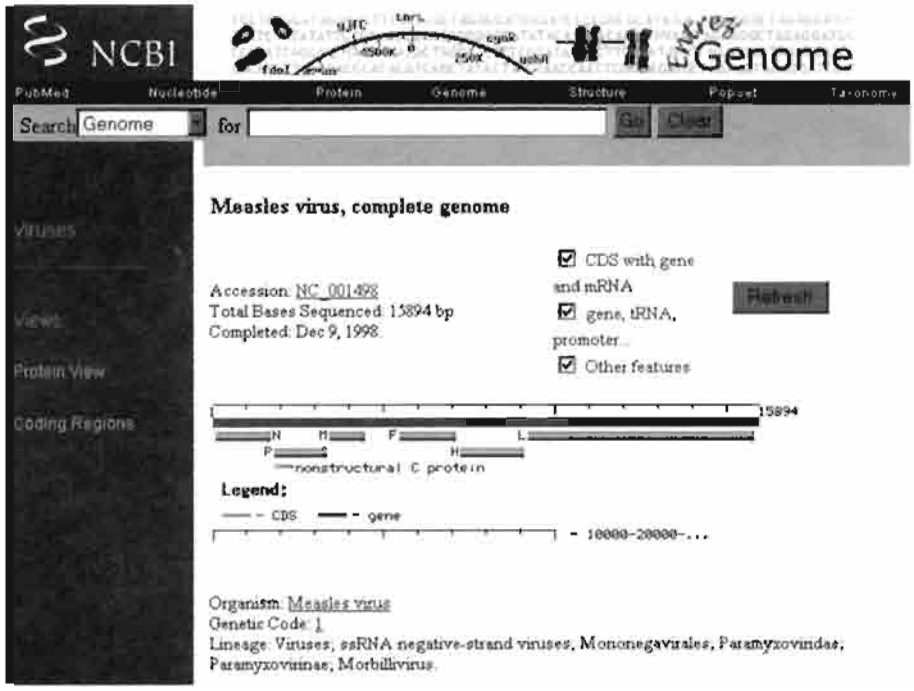
The Viral Genome Organizer was developed by Chris Upton of the University of Victoria (Canada) (▸ http://athena.bioc.uvic.ca/pbr/vgo/).

Genome Organizer, for the graphical display of viral sequences (Upton et al., 2000) (Fig. 13.19). This site also contains a Viral Genome DataBase (VGDB) with analyses of the properties of viral genomes such as GC content. The Stanford HIV RT and Protease Sequence Database offers an algorithm that can be queried with an input viral DNA sequence (Rhee et al., 2003) (Fig. 13.20). The output describes possible mutations in the viral gene and an interpretation of likely susceptibility of that protein to drug resistance.

PERSPECTIVES

Several thousand species of viruses are known. In contrast, there may be tens or hundreds of millions of species of bacteria and archaea (Chapter 14) and perhaps tens of millions of eukaryotic species (Chapters 15–16). There are probably relatively few species of viruses because of their specialized requirements for replication in host cells.

(*a*)



(*b*)

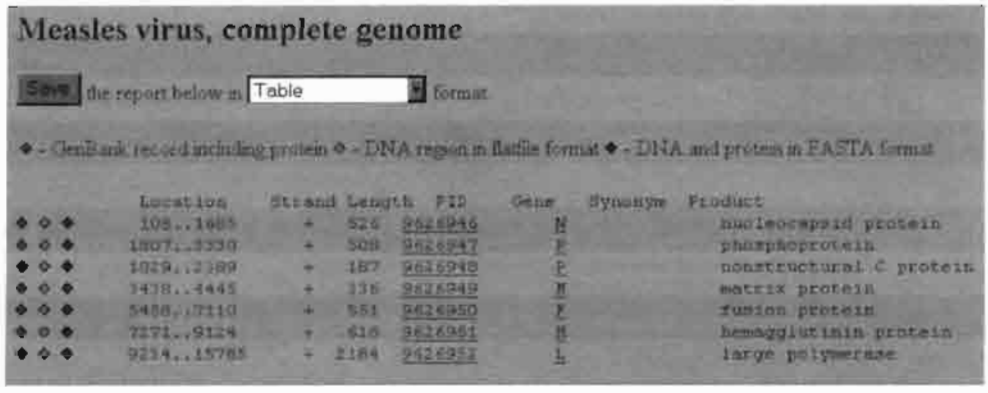


FIGURE 13.17. *Analysis of the measles virus genome.* (a) *The measles virus entry is accessed from the NCBI Entrez website. By clicking on the "Coding Regions" option on the left sidebar,* (b) *a list of the protein-coding genes is obtained. These genes are designated N (nucleocapsid), P (phosphoprotein), M (matrix), F (fusion), H (hemagglutinin), and L (large polymerase). Note that the P gene is predicted to encode another protein (non-structural C protein) using an alternative start site on a different reading frame.*

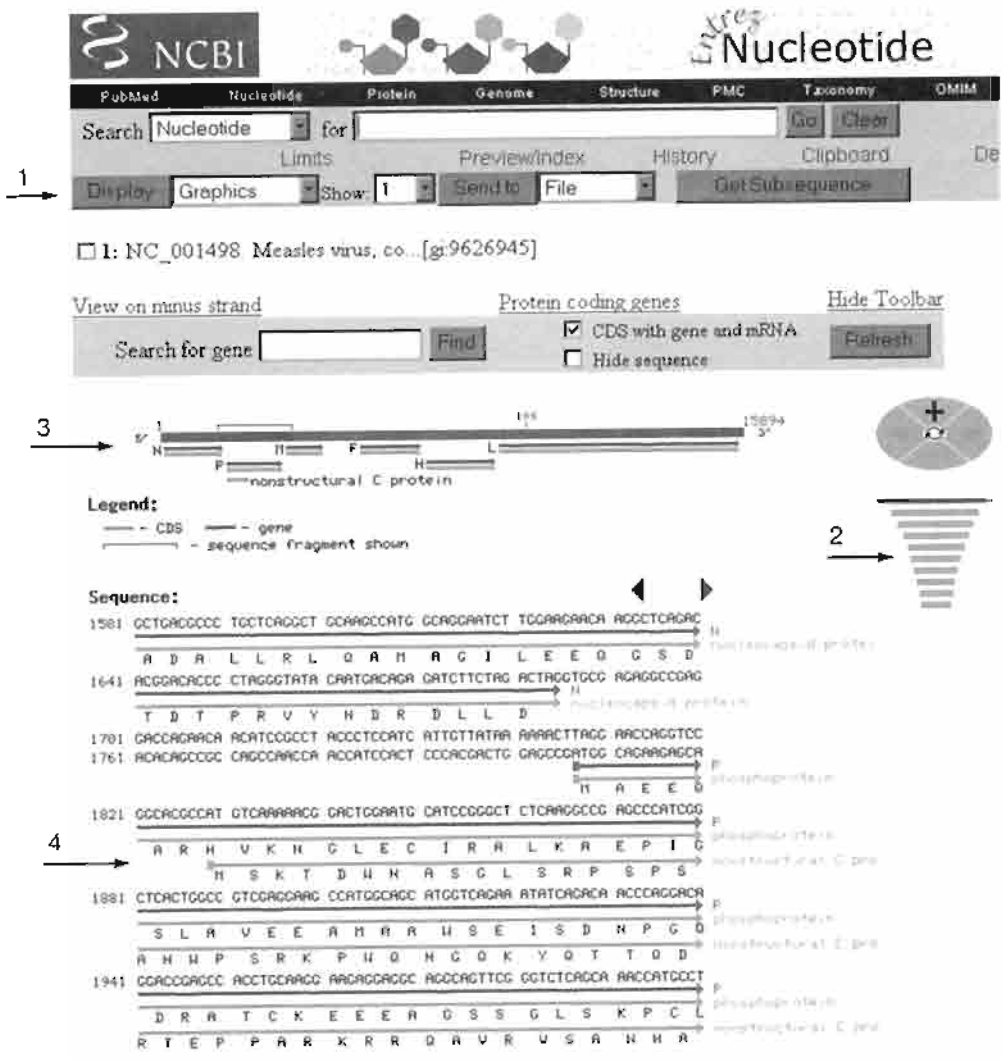


FIGURE 13.18. *The "graphics" display (arrow 1) of any Entrez nucleotide entry shows a map of the DNA sequence along with the corresponding protein sequences. It is easy to zoom in or out for a more detailed or global view of the sequence (arrow 2). The portion of the measles genome that is displayed is indicated in a graphical overview (arrow 3). Usage of an alternative start site allows the measles genome to encode two distinct proteins using nonoverlapping reading frames (arrow 4).*

Essentially all the bioinformatic tools that are applied to eukaryotic or prokaryotic protein and nucleic acid sequences are applicable to the study of viruses as well (Kellam, 2001).

- BLAST, PSI-BLAST, and other database searches may be applied to define the homology of viral sequences to other molecules.
- Microarrays have been used to represent viral genes, allowing an assessment of viral gene transcription during different phases of the viral life cycle.
- In independent approaches, the transcriptional response of host cells to viral infection has begun to be characterized.
- Structural genomics approaches to viruses result in the identification of three-dimensional structures of viral proteins. Some structures are solved in the presence of pharmacological inhibitors. The Entrez protein division of NCBI currently includes over 1700 virus structural records.

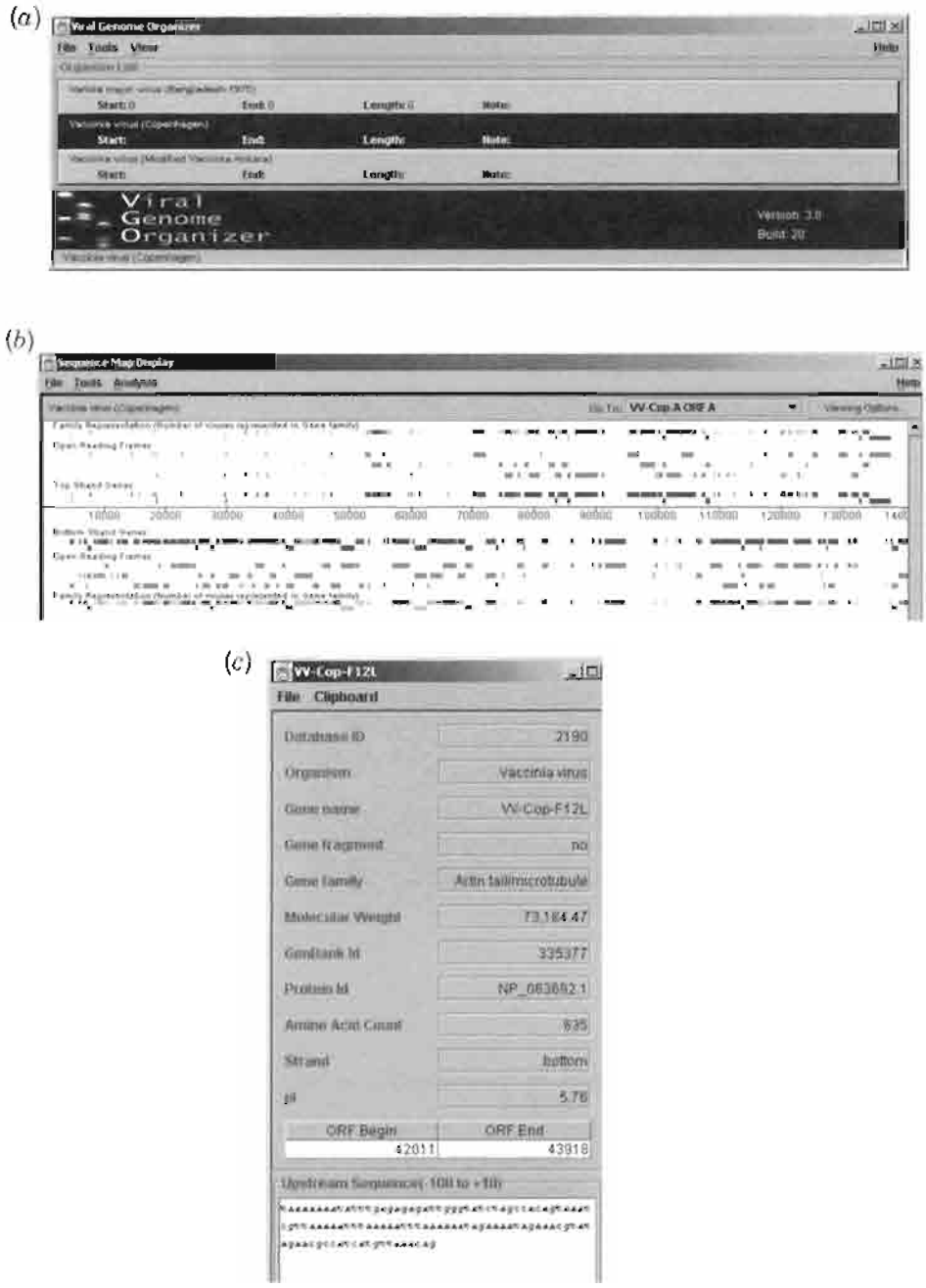


FIGURE 13.19. (a) *The Viral Genome Organizer is a Java-based utility for the analysis of viral genomes. Several dozen complete genome sequences (mostly large poxviruses) can be uploaded.* (b) *The annotated sequences are displayed.* (c) *Each gene is clickable with additional, detailed annotation data.*

PITFALLS

Viruses evolve extremely rapidly, in large part because some viral polymerases tend to operate with low fidelity. It is for this reason that a person infected with HIV may harbor millions of distinct forms of the virus, each with its own unique RNA sequence. Thus, it may be difficult to define a single canonical sequence for some viruses. This complicates attempts to study the evolution of viruses and the functions of their genes.

While the tree of life has been described using rRNA or other sequences (Chapter 12), viruses are almost entirely absent from this tree. This is because there are no genes or proteins that all viruses share in common with other life forms or with each other.

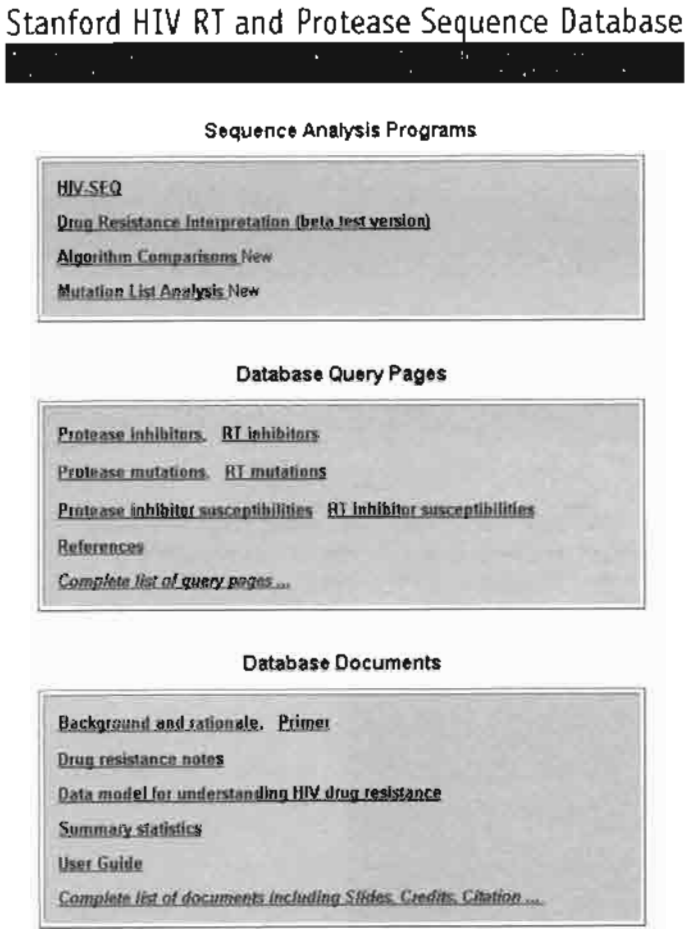


FIGURE 13.20. *Stanford HIV RT database (▸ http://hivdb.stanford.edu/hiv/) focuses on detailed analyses of reverse transcriptase and protease sequences.*

WEB RESOURCES

TABLE 13-3 Virus Resources Available on Web

Resource	Description	URL
ICTVdB	Universal virus database	▸ http://www.ncbi.nlm.nih.gov/ICTVdb/Ictv/fr-index.htm
All the Virology on the WWW	Provides many virology links and resources	▸ http://www.tulane.edu/~dmsander/garryfavweb.html
The Big Picture Book of Viruses	General virus resource	▸ http://www.virology.net/Big_Virology/BVHomePage.html
VIrus Particle ExploreER (VIPER)	High-resolution virus structures in the Protein Data Bank (PDB)	▸ http://mmtsb.scripps.edu/viper/
Project VirgO	Software tools for the analysis of viral genomes	▸ http://athena.bioc.uvic.ca/genomes/index.html
Viral Genome Organizer	Analyses of large poxviruses and other viruses	▸ http://athena.bioc.uvic.ca/pbr/vgo/
The EMBL Virus Structure Resource	Provides images of virus structures	▸ http://www.embl-heidelberg.de/ExternalInfo/fuller/icos0.html
Institute for Molecular Virology	A research institute at the University of Wisconsin-Madison	▸ http://virology.wisc.edu/IMV/about.html
Taxonomy of Plant Viruses	Virus classification	▸ http://www.scri.sari.ac.uk/TiPP/ICTV/ictvhome.html
Stanford HIV RT and Protease Sequence Database	A curated database with information on drug targets	▸ http://hivdb.stanford.edu/hiv/
INDEX VIRUM	Catalog of viruses	▸ http://life.anu.edu.au/viruses/Ictv/fr-index.htm

Discussion Questions

[13-1] There is no comprehensive molecular phylogenetic tree of all viruses. Why not?

[13-2] If you wanted to generate phylogenetic trees that are as comprehensive as possible, using DNA or RNA or protein sequences available in GenBank, what molecule(s) would you select?

Problems

[13-1] How many HIV-1 proteins are in LocusLink? Given the tremendous heterogeneity of HIV-1, you might expect there to be thousands of variant forms of each protein. How many are actually assigned RefSeq accession numbers?

[13-2] Find an HIV-1 protein with a RefSeq identifier in LocusLink (such as the Vif protein, NP_057851; you should select your own example). Perform a blastp search with it, and inspect the results using the taxonomy report. Next, repeat the search, excluding HIV from the output. As an example of how to do this, enter "vif NOT txid11676[Organism]" or "vif NOT Hiv[Organism]" into the advanced search option "Limit by Entrez query." (Note that you can find the taxonomy identifier txid11676[Organism] by using the NCBI taxonomy browser.) How broadly is the gene or protein you selected represented among viruses? Do you expect some genes to be HIV specific while other genes are shared broadly by viruses?

Self-Test Quiz

[13-1] There are several thousand known viruses, while there are many millions of prokaryotes and eukaryotes. The most likely explanation for the small number of viruses is that
- (a) we have not yet learned how to detect most viruses
- (b) we have not yet learned how to sequence most viruses
- (c) there are few viruses because their needs for survival are highly specialized
- (d) viruses use an alternative genetic code

[13-2] The HIV genome contains nine protein-coding genes. The number of GenBank accession numbers for these nine genes is approximately
- (a) 9
- (b) 900
- (c) 9,000
- (d) 90,000

[13-3] For functional genomics analyses of viruses, it is possible to measure gene expression
- (a) of viral genes upon viral infection of human tissues
- (b) of human genes upon viral infection of human tissues
- (c) of viral genes and human genes, simultaneously measured upon viral infection of human tissue
- (d) of viral genes or human genes, separately measured upon viral infection of human tissue

[13-4] Herpesviruses probably first appeared about
- (a) 200 million years ago
- (b) 2 million years aog
- (c) 20,000 years ago
- (d) 200 years ago

[13-5] HIV probably first appeared about
- (a) 70 million years ago
- (b) 7 million years ago
- (c) 7,000 years ago
- (d) 70 years ago

[13-6] Phylogeny of HIV virus subtypes
- (a) establishes that HIV emerged from a cattle virus
- (b) can be used to develop vaccines directed against ancestral protein sequences
- (c) establishes which human tissues are most susceptible to infection

[13-7] Specialized virus databases such as that at Oak Ridge National Laboratory offer resources for the study of HIV that are not available at NCBI or EBI. An example is:
- (a) a listing of thousands of variant forms for each HIV gene
- (b) a listing of literature and citations from the previous week
- (c) graphical displays of the genome
- (d) a description of where HIV variants have been identified across the world

Suggested Reading

Kellam (2001) has reviewed bioinformatics and functional genomics approaches to virology. He describes a broad range of subjects from viral genome structure to protein structure and gene expression studies of both viral pathogens and hosts.

Flint et al. (2000) edited an authoritative textbook on virology. Schaechter et al. (1999) edited a book on the mechanisms of microbial disease. They include chapters on a broad range of viruses presented from both a basic science and a clinical perspective.

REFERENCES

Ackermann, H. W. Frequency of morphological phage descriptions in the year 2000. *Arch. Virol.* **146,** 843–857 (2001).

Barrett, T., and Rossiter, P. B. Rinderpest: The disease and its impact on humans and animals. *Adv. Virus Res.* **53,** 89–110 (1999).

Büchen-Osmond, C. Further progress in ICTVdB, a universal virus database. *Arch. Virol.* **142,** 1734–1739 (1997).

Chang, Y., et al. Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. *Science* **266,** 1865–1869 (1994).

Cieslak, T. J., Christopher, G. W., and Ottolini, M. G. Biological warfare and the skin II: Viruses. *Clin. Dermatol.* **20,** 355–364 (2002).

Davison, A. J. Evolution of the herpesviruses. *Vet. Microbiol.* **86,** 69–88 (2002).

Flint, S. J., Enquist, L. W., Krug, R. M., Racaniello, V. R., and Skalka, A. M. *Principles of Virology Molecular Biology, Pathogenesis, and Control.* American Society for Microbiology Press, Washington, DC, 2000.

Flores, R. A naked plant-specific RNA ten-fold smaller than the smallest known viral RNA: The viroid. *C. R. Acad. Sci. III* **324,** 943–952 (2001).

Gaschen, B., Kuiken, C., Korber, B., and Foley, B. Retrieval and on-the-fly alignment of sequence fragments from the HIV database. *Bioinformatics* **17,** 415–418 (2001).

Gaschen, B., et al. Diversity considerations in HIV-1 vaccine selection. *Science* **296,** 2354–2360 (2002).

Hahn, B. H., Shaw, G. M., De Cock, K. M., and Sharp, P. M. AIDS as a zoonosis: Scientific and public health implications. *Science* **287,** 607–614 (2000).

Hendrix, R. W. Evolution: The long evolutionary reach of viruses. *Curr. Biol.* **9,** R914–917 (1999).

Johnson, R. T., and Gibbs, C. J., Jr. Creutzfeldt–Jakob disease and related transmissible spongiform encephalopathies. *N. Engl. J. Med.* **339,** 1994–2004 (1998).

Johnson, C. E., et al. Measles vaccine immunogenicity and antibody persistence in 12 vs 15-month old infants. *Vaccine* **18,** 2411–2415 (2000).

Kellam, P. Post-genomic virology: The impact of bioinformatics, microarrays and proteomics on investigating host and pathogen interactions. *Rev. Med. Virol.* **11,** 313–329 (2001).

La Scola, B., et al. A giant virus in Amoebae. *Science* 299, 2033 (2003).

Mayo, M. A., and Pringle, C. R. Virus taxonomy–1997. *J. Gen. Virol.* **79,** 649–657 (1998).

McClure, M. A. The complexities of genome analysis, the Retroid agent perspective. *Bioinformatics* **16,** 79–95 (2000).

McGeoch, D. J., Cook, S., Dolan, A., Jamieson, F. E., and Telford, E. A. Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. *J. Mol. Biol.* **247,** 443–458 (1995).

Meissner, C., and Coffin, J. M. The human retroviruses: AIDS and other diseases. In M. Schaechter, N. C. Engleberg, B. I. Eisenstein, and G. Medoff (Eds.), *Mechanisms of Microbial Disease.* Lippincott Williams & Wilkins, Baltimore, MD, 1999, Chapter 38.

Paulose-Murphy, M., et al. Transcription program of human herpesvirus 8 (kaposi's sarcoma-associated herpesvirus). *J. Virol.* **75,** 4843–4853 (2001).

Poole, L. J., et al. Altered patterns of cellular gene expression in dermal microvascular endothelial cells infected with Kaposi's sarcoma-associated herpesvirus. *J. Virol.* **76,** 3395–3420 (2002).

Prusiner, S. B. Prions. *Proc. Natl. Acad. Sci. USA* **95,** 13363–13383 (1998).

Reid, A. H., Fanning, T. G., Janczewski, T. A., McCall, S., and Taubenberger, J. K. Characterization of the 1918 "Spanish" influenza virus matrix gene segment. *J. Virol.* **76,** 10717–10723 (2002).

Rhee, S. Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J., and Shafer, R. W. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **31,** 298–303 (2003).

Russo, J. J., et al. Nucleotide sequence of the Kaposi sarcoma-associated herpesvirus (HHV8). *Proc. Natl. Acad. Sci. USA* **93,** 14862–14867 (1996).

Schaechter, M., Engleberg, N. C., Eisenstein, B. I., and Medoff, G. *Mechanisms of Microbial Disease.* Lippincott Williams & Wilkins, Baltimore, MD, 1999.

Sharp, P. M., et al. The origins of acquired immune deficiency syndrome viruses: Where and when? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **356,** 867–876 (2001).

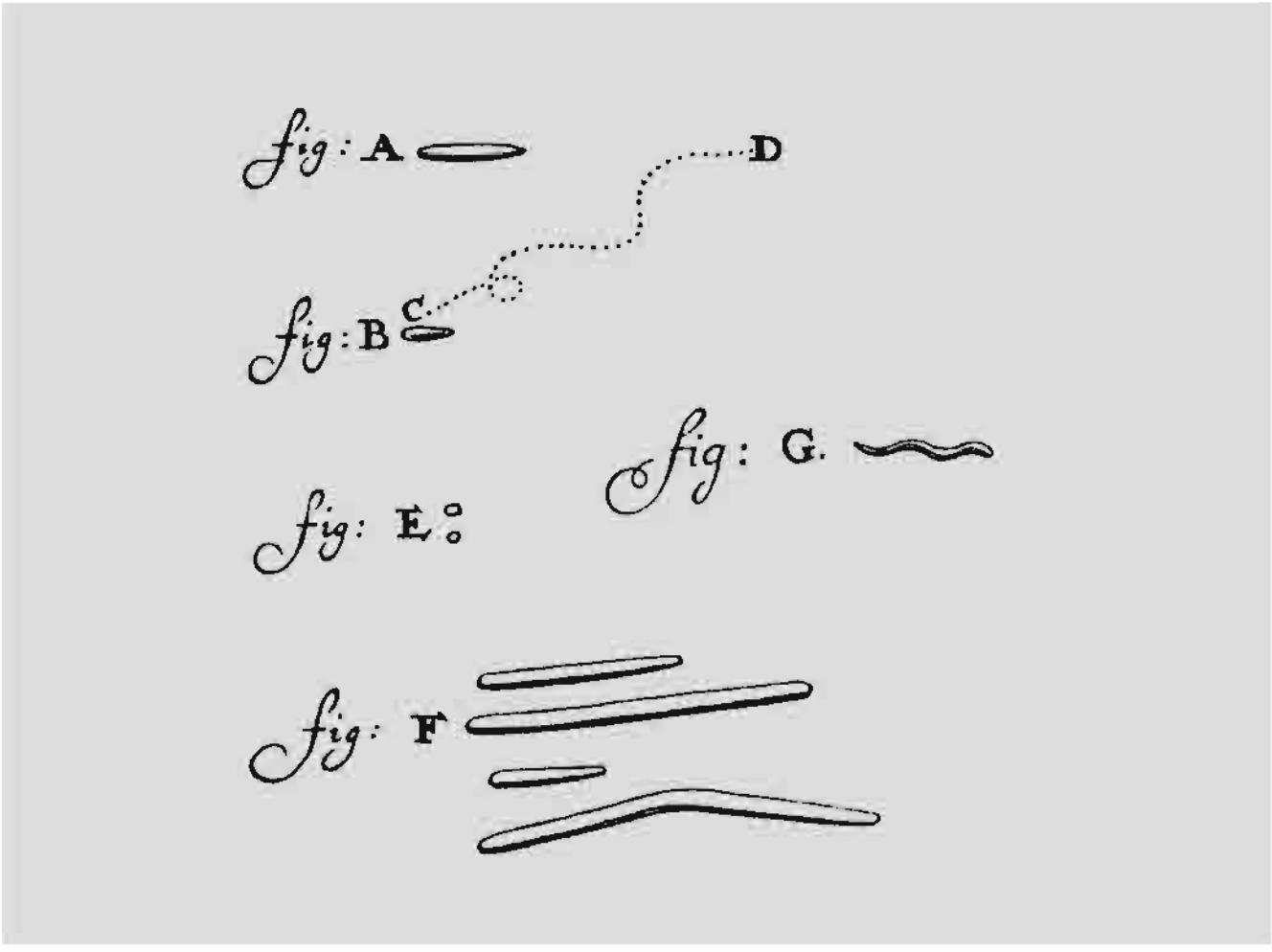
Simpson, G. G. The meaning of taxonomic statements. In S. L. Washburn (ed.). *Classification and Human Evolution.* Aldine Publishing Co., Chicago, 1963, pp. 1–31.

Suarez, D. L., et al. Comparisons of highly virulent H5N1 influenza A viruses isolated from humans and chickens from Hong Kong. *J. Virol.* **72,** 6678–6688 (1998).

Upton, C., Hogg, D., Perrin, D., Boone, M., and Harris, N. L. Viral genome organizer: A system for analyzing complete viral genomes. *Virus. Res.* **70,** 55–64 (2000).

Wakeling, M. N., Roy, D. J., Nash, A. A., and Stewart, J. P. Characterization of the murine gammaherpesvirus 68 ORF74 product: A novel oncogenic G protein-coupled receptor. *J. Gen. Virol.* **82,** 1187–1197 (2001).

This Page Intentionally Left Blank



Antony van Leeuwenhoek (1622–1723) has been called the father of protozoology and bacteriology. This figure shows bacteria he observed taken from his own mouth. Figure A indicates a motile Bacillus. *Figure B shows* Selenomonas sputigena, *while C and D show the path of its motion. Figure E shows two micrococci; F shows* Leptotrichia buccalis, *and G shows a spirochete. He describes these "animalcules," found in his and others' mouths, in a letter written 17 September 1683. "While I was talking to an old man (who leads a sober life, and never drinks brandy or [smokes] tobacco, and very seldom any wine), my eye fell upon his teeth, which were all coated over; so I asked him when he had last cleaned his mouth? And I got for answer that he'd never washed his mouth in all his life. So I took some spittle out of his mouth and examined it; but I could find in it nought but what I had found in my own and other people's. I also took some of the matter that was lodged between and against his teeth, and mixing it with his own spit, and also with fair water (in which there were no animalcules), I found an unbelievably great company of living animalcules, a-swimming more nimbly than any I had ever seen up to this time. The biggest sort (where of there were a great plenty) bent their body into curves in going forwards, as in Fig. G. Moreover, the other animalcules were in such enormous numbers, that all the water (notwithstanding only a very little of the matter taken from between the teeth was mingled with it) seemed to be alive" (translated from the Dutch by Dobell, 1932, pp. 242–243). Used with permission.*

14

Completed Genomes: Bacteria and Archaea

"And now you may be disposed to ask: To what end is this discourse on the anatomy of beings too minute for ordinary vision, and of whose very existence we should be ignorant unless it were revealed to us by a powerful microscope? What part in nature can such apparently insignificant animalcules play, that can in any way interest us in their organization, or repay us for the pains of acquiring a knowledge of it? I shall endeavour briefly to answer these questions. The Polygastric Infusoria, notwithstanding their extreme minuteness, take a great share in important offices of the economy of nature, on which our own well-being more or less immediately depends.

Consider their incredible numbers, their universal distribution, their insatiable voracity; and that it is the particles of decaying vegetable and animal bodies which they are appointed to devour and assimilate.

Surely we must in some degree be indebted to those ever active invisible scavengers for the salubrity of our atmosphere. Nor is this all: they perform a still more important office, in preventing the gradual diminution of the present amount of organized matter upon the earth. For when this matter is dissolved or suspended in water, in that state of comminution and decay which immediately precedes its final decomposition into the elementary gases, and its consequent return from the organic to the inorganic world, these wakeful members of nature's invisible police are every where ready to arrest the fugitive organized particles, and turn them back into the ascending stream of animal life."

—Richard Owen (1843, p. 27)

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

INTRODUCTION

In this chapter we will consider bioinformatic approaches to two of the three main branches of life: bacteria and archaea. Bacteria and archaea are grouped together because they are prokaryotes, that is, single-celled organisms that lack nuclei. Bacteria and archaea are also termed microorganisms. The term microbe refers to those microorganisms that cause disease in humans; microbes include many eukaryotes such as fungi and protozoa (Chapters 15 and 16) as well as some prokaryotes.

It has been estimated that there are 10^{30} bacteria, comprising the majority of the biomass on the planet (Sherratt, 2001).

It has been estimated that bacteria account for 60% of Earth's biomass. Bacteria occupy every conceivable ecological niche in the planet, and there may be as many as 2–3 billion distinct bacterial species (Fraser et al., 2000). The great majority of bacteria and archaea (>99%) have never been cultured or characterized (DeLong and Pace, 2001). A compelling reason to study bacteria is that many cause disease in humans and other animals.

This chapter provides an overview of bioinformatic approaches to the study of bacteria and archaea. We review aspects of prokaryotic biology such as genome size and complexity, and tools for the analysis and comparison of prokaryotic genomes.

In Chapter 12, we described many of the genome-sequencing projects for bacteria and archaea in chronological order, beginning with the sequencing of *Haemophilus influenzae* in 1995. We will now consider the classification of bacteria and archaea by six different criteria: (1) morphology, (2) genome size, (3) lifestyle, (4) relevance to human disease, (5) molecular phylogeny using rRNA, and (6) molecular phylogeny using other molecules. We will introduce several key resources for the study of bacterial and archaeal genomes: NCBI and the Comprehensive Microbial Resource (CMR) at The Institute for Genomic Research (TIGR). In describing these resources, we will examine bioinformatics tools to analyze individual microbial genomes and tools for the comparison of two or more genomes. It is through comparative genomics that we are beginning to appreciate some of the important principles of microbial biology, such as the adaptation of microbes to highly specific ecological niches, the lateral transfer of genes between microbes, genome expansion and reduction, and the molecular basis of pathogenicity.

Pathogenicity is the ability of an organism to cause disease. Virulence is the degree of pathogenicity.

CLASSIFICATION OF BACTERIA BY MORPHOLOGICAL CRITERIA

Most bacteria are classified into four main types: gram-positive and gram-negative cocci or rods (reviewed in Schaechter, 1999). Examples of these different bacteria are presented in Table 14.1. The Gram stain is absorbed by about half of all bacteria and reflects the protein and peptidoglycan composition of the cell wall. Many other bacteria do not fit the categories of gram-positive or gram-negative cocci or rods because they have atypical shapes or staining patterns. As an example, spirochetes such as the Lyme disease agent *Borrelia burgdorferi* have a characteristic outer membrane sheath, protoplasmic cell cylinder, and periplasmic flagella (Charon and Goldstein, 2002).

The classification of microbes based on molecular phylogeny is far more comprehensive, as described below. Molecular differences can reveal the extent of microbial diversity both between species (showing the breadth of the prokaryotic tree of life) and within species (e.g., showing molecular differences in pathogenic isolates and in closely related, nonvirulent strains). However, beyond molecular

TABLE 14-1 Major Categories of Bacteria Based on Morphological Criteria

Type	Examples[a]
Gram-positive cocci	*Streptococcus pyogenes, Staphylococcus aureus*
Gram-positive rods	*Corynebacterium diphtheriae, Bacillus anthracis* (anthrax), *Clostriduium botulinum*
Gram-negative cocci	*Neisseria, Gonococcus*
Gram-negative rods	*Escherichia coli, Vibrio cholerae, Helicobacter pylori*
Other	*Mycobacterium leprae* (leprosy), *Borrelia burgdorferi* (Lyme disease), *Chlamydia trachomatis* (sexually transmitted disease), *Mycoplasma pneumoniae*

[a] The disease is indicated in parentheses.

criteria there are many additional ways to differentiate bacteria based on microscopy and studies of physiology—for example, distinguishing those microbes that are capable of oxygenic photosynthesis (Cyanobacteria) or those that produce methane.

CLASSIFICATION OF BACTERIA AND ARCHAEA BASED ON GENOME SIZE AND GEOMETRY

In haploid organisms such as bacteria and archaea, the genome size (or C value) is the total amount of DNA in the genome. Bacterial and archaeal genomes vary in size over about a 25-fold range from about 500,000 bp [0.5 megabases (Mb)] to over 10 Mb (Table 14.2) (Casjens, 1998). The genome sizes of 23 named major bacterial phyla and some of their subgroups are shown in Figure 14.1. As indicated in the figure, most bacterial genomes are circular, although some are linear; some bacterial genomes consist of multiple circular chromosomes. Plasmids (small circular extrachromosomal elements) have been found in most bacterial phyla, although linear extrachromosomal elements are more rare.

In diploid or polyploid organisms, the genome size is the amount of DNA in the unreplicated haploid genome (such as the sperm cell nucleus). We will discuss eukaryotic genome sizes in Chapters 15–17.

The genome of the fission yeast *Schizosaccharomyces pombe* is 13 Mb in size and encodes about 4800 genes (Wood et al., 2002) (Chapter 15). Thus some bacterial genomes are considerably larger than eukaryotic genomes. The *Streptomyces coelicolor* genome, the largest bacterial genome that has been sequenced to date, is 8.7 Mb and encodes over 7800 proteins (Bentley et al., 2002).

Overall, the number of genes encoded in a bacterial genome ranges from about 500 to 8000. This is a 16-fold range, comparable to the range in C values. For a

TABLE 14-2 Range of Genome Sizes in Bacteria and Archaea

Taxon	Genome Size Range (Mb)	Ratio (Highest/Lowest)
Bacteria	0.58–13.2	23
Mollicutes	0.58–2.2	4
Gram negative	0.65–9.5	15
Gram positive	1.6–11.6	7
Cyanobacteria	3.1–13.2	4
Archaea	0.5–4.1	3

Source: Modified from Graur and Li (2000, p. 36). Used with permission.

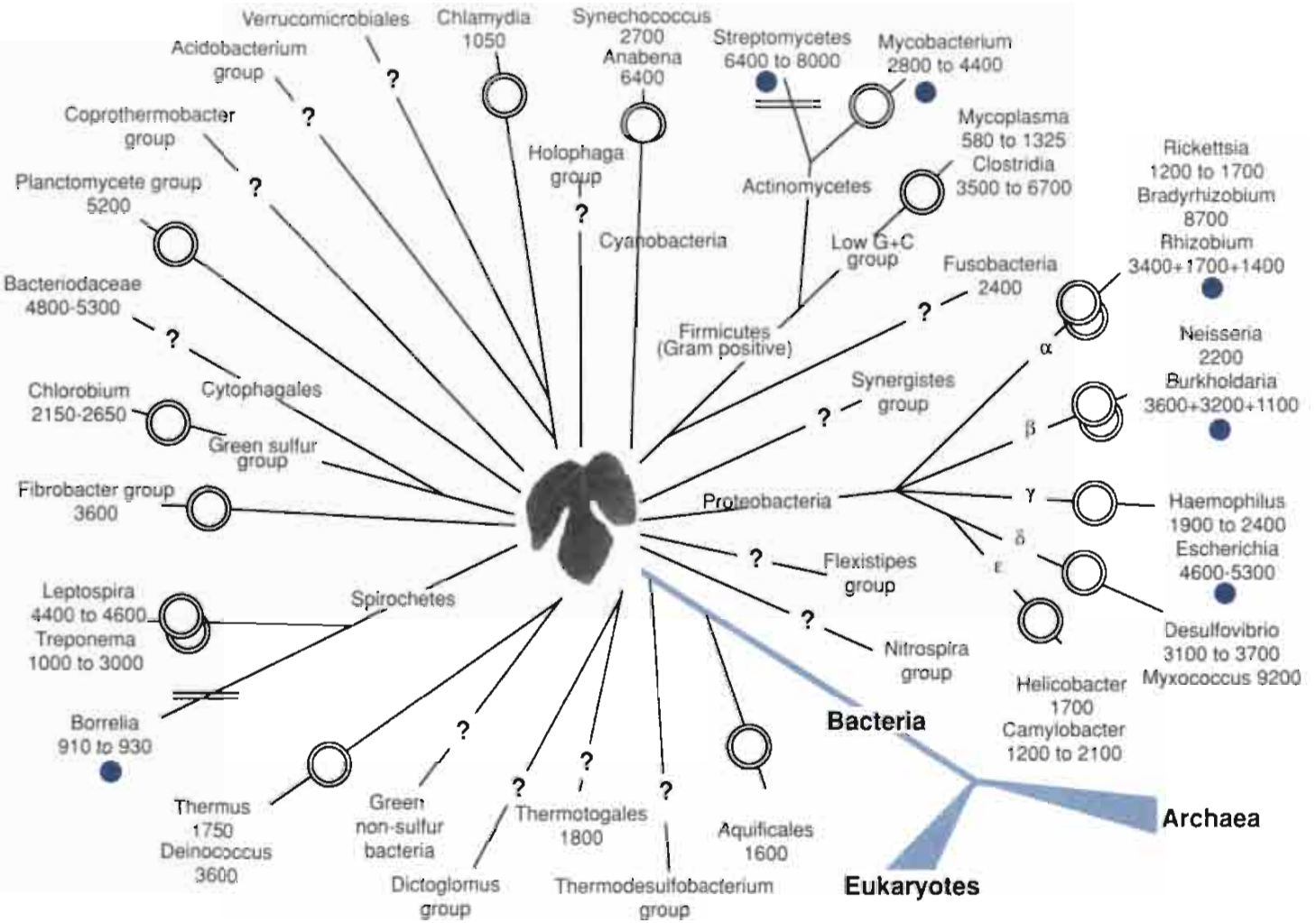


FIGURE 14.1. *Bacterial chromosome size and geometry. The 23 named major bacterial phyla are represented as well as some of their subgroups. The tree is based on rRNA sequences and is unrooted. The branch lengths do not depict phylogenetic distances, and the fig leaf at the center indicates uncertain branching patterns. The chromosome geometry (circular or linear, in some cases with multiple chromosomes) is indicated at the end of each branch. The chromosome sizes of representative genera are given (in kilobases). Linear extrachromosomal elements, common in borrelias and actinomycetes, are indicated. This figure was modified from Casjens (1998). Used with permission.*

TABLE 14-3 Genome Size of Selected Bacteria and Archaea Having Relatively Large or Small Genomes

Species	Genome Size (Mb)	Coding Regions	GC Content	Reference
Streptomyces coelicolor	8.67	7825	72	Bentley et al., 2002
Burkholderia cepacia	7.6 (3.7 + 3.1 + 0.8)		65	Unpublished
Pseudomonas fluorescens	6.6	NA	66	Unpublished
Mycobacterium marinum	6.0	NA	65	Unpublished
Tropheryma whippelii	0.93	808	46	Unpublished
Ureaplasma urealyticum parvum biovar serovar 3	0.752	613	26	Glass et al., 2000
Mycoplasma pneumoniae M129	0.816	677	40	Himmelreich et al., 1996
Mycoplasma genitalium G-37	0.58	470	32	Fraser et al., 1995
Nanarchaeum equitans	0.5	NA	NA	Huber et al., 2002

Source: Adapted from ▸ http://www.sanger.ac.uk/Projects/Microbes/ and the NCBI website (PubMed, Entrez). NA, not available.

number of bacteria with completely sequenced genomes, protein-coding genes constitute about 85–94% of the genome. Thus, intergenic and nongenic fractions are small. [An exception is the pathogen that causes leprosy, *Mycobacterium leprae.* Its genome underwent massive gene decay, and protein-coding genes constitute only 49.5% of the genome (Cole et al., 2001).] The density of genes in microbial genomes is consistently about one gene per kilobase. Even in very small genomes such as *Mycoplasma genitalium,* reduced genome sizes are not associated with changes either in gene density or in the average size of genes (Fraser et al., 1995). The genome sizes of selected large or small bacteria and archaea are shown in Table 14.3.

Another exception is the parasite *Rickettsia prowazekii,* described below, that has 24% noncoding DNA.

The smallest sequenced genome of a free-living organism is that of *Mycoplasma genitalium,* a urogenital pathogen. The *M. genitalium* has 580,070 bp encoding 470 protein-coding genes, 3 rRNA genes, and 33 tRNA genes (Fraser et al., 1995). Mycoplasmas are bacteria of the class Mollicutes. They lack a cell wall and have a low GC content (32%) characteristic of this class.

We introduced ways to view information about any bacterial genome, such as *M. genitalium,* at the NCBI website (Chapter 12). Another important website is CMR at TIGR. CMR contains robust annotation of completed microbial genomes (Peterson et al., 2001). The entry for *M. genitalium* is shown in Figure 14.2, including links to PubMed, background information on the biology of the organism, file transfer protocol sites to download the nucleotide or protein sequences, and a display of the chromosome. This display includes a clickable map with genome annotations (Fig. 14.3). The CMR offers dozens of tools for prokaryotic genome analysis that we will explore in this chapter.

To access CMR, go to TIGR (▸ http://www.tigr.org) and click the CMR link.

An even smaller genome than *M. genitalium* has been discovered in a hyperthermophilic archaeon that was cultured from a submarine hot vent, *Nanoarchaeum equitans* (Huber et al., 2002). This organism appears to grow attached to another archeon, *Ignicoccus.* Because of its small cell size (400 nm) and small genome size, Huber et al. (2002) suggested that *N. equitans* resembles an intermediate between the smallest living organisms (such as *M. genitalium*) and large viruses (such as the pox virus). Nonetheless, even parasitic intracellular bacteria and archaea are classified as free-living organisms distinct from viruses. By comparing small prokaryotic genomes, it is possible to estimate the minimal number of genes required for life (Box 14.1).

FIGURE 14.2. *The CMR at The Institute for Genomic Research provides one of the most important websites for the study of microbes (▶ http://www.tigr.org). The page for a typical organism is displayed, including a clear taxonomic description, links to literature and background information, and access to the annotated genome sequence.*

CLASSIFICATION OF BACTERIA AND ARCHAEA BASED ON LIFESTYLE

This classification of bacteria is derived in part from an online essay by Jim Moulder (▶ http://www.chlamydiae.com/Evolution_index.htm).

In addition to the criteria of morphology and genome size and geometry, a third approach to classifying bacteria (and archaea) is based on their lifestyle. One main advantage of this approach to classifying microbes is that it conveniently highlights the principle of extreme reduction in genome size that is associated with three lifestyles: extremophiles and intracellular and epicellular prokaryotes:

- Extremophiles are microbes that live in extreme environments. Archaea have been identified in hypersaline conditions (halophilic archaea), geothermal

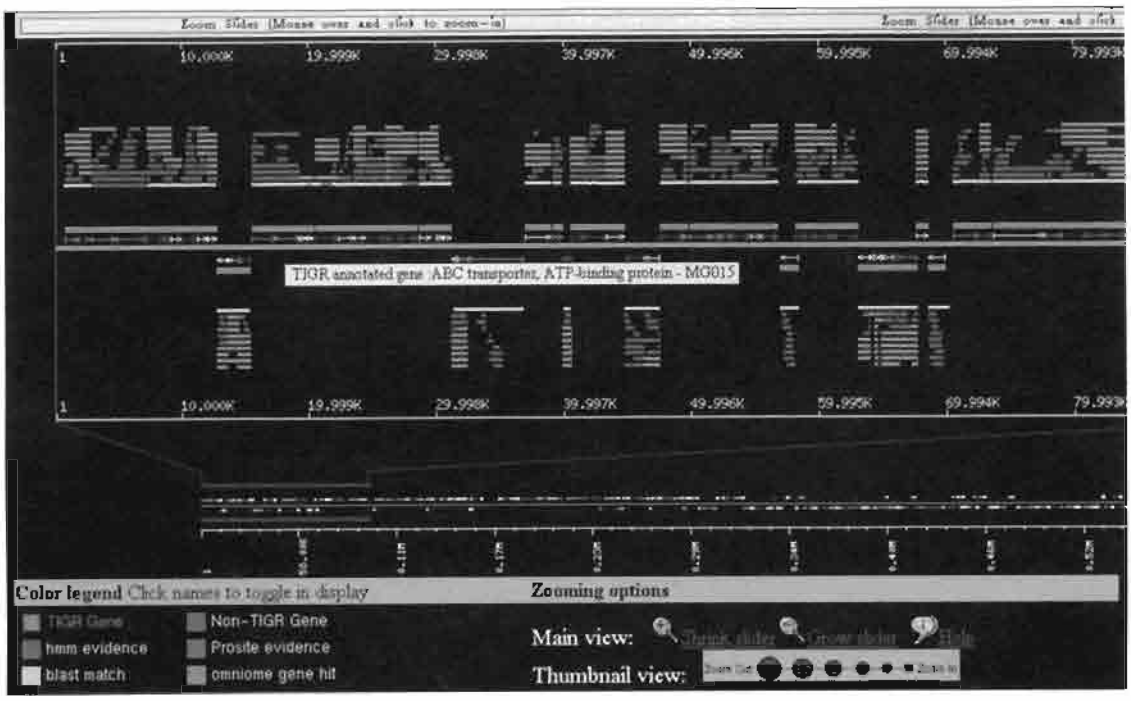


FIGURE 14.3. By clicking on the genome graphic from a TIGR CMR entry (Fig. 14.2), you can access an annotated genome viewer for any of the dozens of genomes in the database.

areas such as hot vents (hyperthermophilic archaea), and anoxic habitats (methanogens) (DeLong and Pace, 2001).

- Intracellular bacteria invade eukaryotic cells; a well-known example is the α-proteobacterium that is thought to have invaded eukaryotic cells and evolved into the present-day mitochondrion.
- Epicellular bacteria (and archaea) are parasites that live in close proximity to their hosts but not inside host cells.

BOX 14-1
Smallest Genome size: What Is the Minimal Genome Size for Life?

How many genes are required in the genome of the smallest living organism—that is, the smallest autonomous self-replicating organism? An approach to this question is to identify the orthologs in common between several microbes. Mushegian and Koonin (1996) identified 239 genes in common between *Escherichia coli, H. influenzae,* and *M. genitalium*. This is considered an estimate of the minimal genome size. However, nonorthologous gene displacement may have occurred in which genes that are essential for life are functionally related but not related at the sequence level. Of the 239 genes identified by Mushegian and Koonin, the functions include several basic categories: translation, DNA replication, recombination and DNA repair, transcription, anaerobic metabolism, lipid and cofactor biosynthesis, transmembrane transporters. They also identified many proteins of unknown function.

Notably, *H. influenzae* (a gram-negative bacterium) and *M. genitalium* (a gram-positive bacterium) are extremely distantly related, having shared a common ancestor 2 BYA (Feng et al., 1997). Thus Mushegian and Koonin's list of minimal gene content is unlikely to reflect the gene content of the common ancestor of life (Doolittle, 1998).

An experimental approach to determining the minimal number of genes required for life was taken by Itaya (1995), who randomly knocked out protein-coding genes in the bacterium *Bacillus subtilis*. Mutations in only 6 of 79 loci prevented growth of the bacteria and were indispensible. Extrapolating to the size of the complete *B. subtilis* genome, about 250 genes are estimated to be essential for life.

TABLE 14-4 Classification of Bacteria and Archaea Based on Ecological Niche

Lifestyle	Bacterium	Genome Size (Mb)	Reference
Extracellular	*Escherichia coli*	4.6	Blattner et al., 1997
	Vibrio cholerae	4.0	Heidelberg et al., 2000
	Pseudomonas aeruginosa	6.3	Stover et al., 2000
	Bacillus subtilis	4.2	Kunst et al., 1997
	Clostridium acetobutylicum	4.0	Nolling et al., 2001
	Deinococcus radiodurans	3.3	White et al., 1999
Facultatively intracellular	*Salmonella enterica*	4.8	Parkhill et al., 2001a
	Yersinia pestis	4.7	Parkhill et al., 2001b
	Legionella pneumophila	3.9	Bender et al., 1990
	Mycobacterium tuberculosis	4.4	Cole et al., 1998
	Listeria monocytogenes	2.9	Glaser et al., 2001
Extremophile	*Aeropyrum pernix*	1.7	Kawarabayasi et al., 1999
	Methanococcus janneschi	1.7	Bult et al., 1996
	Archeoglobus fulgidus	2.2	Klenk et al., 1997
	Thermotoga maritima	1.9	Nelson et al., 1999
	Aquifex aeolius	1.6	Deckert et al., 1998
Epicellular	*Neisseria meningitidis*	2.2	Tettelin et al., 2000
	Haemophilus influenzae	1.8	Fleischmann et al., 1995
	Mycoplasma genitalium	0.6	Fraser et al., 1995
	Mycoplasma pneumoniae	0.8	Himmelreich et al., 1996
	Ureaplasma urealyticum	0.8	Glass et al., 2000
	Mycoplasma pulmonis	1.0	Chambaud et al., 2001
	Borrelia burgdorferi	0.9	Fraser et al., 1997; Casjens et al., 2000
	Treponema pallidum	1.1	Fraser et al., 1998
	Helicobacter pylori	1.7	Tomb et al., 1997; Alm et al., 1999
	Pasteurella multocida	2.3	May et al., 2001
Obligate intracellular, symbiotic	*Buchnera* sp.	0.6	Shigenobu et al., 2000
	Wolbachia spp.	1.1	Sun et al., 2001
	Wigglesworthia glossinidia	0.7	Akman et al., 2002
	Sodalis glossinidius	2.0	Akman et al., 2001
Obligate intracellular, parasitic	*Rickettsia prowazekii*	1.1	Andersson et al., 1998
	Rickettsia conorii	1.3	Ogata et al., 2001
	Ehrlichia chaffeensis	1.2	–
	Cowdria ruminantium	1.6	de Villiers et al., 2000
	Chlamydia trachomatis	1.1	Stephens et al., 1998; Read et al., 2000
	Chlamydophila pneumoniae	1.3	Kalman et al., 1999; Read et al., 2000; Shirai et al., 2000

Source: Adapted from ▸http://www.chlamydiae.com.

We may distinguish six basic lifestyles of bacteria and archaea (Table 14.4):

Each year, 1.9 million people die of tuberculosis and 1.9 billion people are infected worldwide (▸http://www.cdc.gov/ncidod/eid/vol8no11/02-0468.htm). The *M. tuberculosis* genome was sequenced by Cole et al. (1998).

1. Extracellular: For example, *E. coli* commonly inhabits the human intestine without entering cells.

2. Facultatively intracellular bacteria can enter host cells, but this behavior depends on environmental conditions. *Mycobacterium tuberculosis,* the cause of tuberculosis, can remain dormant within infected macrophages, only to activate and cause disease many decades later.

3. Extremophilic microbes: Initially, archaea were all identified in extreme environmental conditions. Some archaea have been found to grow at temperatures as high as 113°C, at pH 0, and in salt concentrations as high as 5 *M* sodium chloride. *Methanococcus janneschi,* the first archeal organism to have its genome completely sequenced (Bult et al., 1996), grows at pressures over 200 atm and at an optimum temperature near 85°C. Archaea have subsequently been identified in less extreme habitats, including forest soil and ocean seawater (DeLong, 1998).

4. Epicellular prokaryotes grow outside of their hosts, but in association with them. *Mycoplasma pneumoniae,* a bacterium with a genome size of ≈816,000 bp, is a major cause of respiratory infections. The bacterium is a surface parasite that attaches to the respiratory epithelium of its host. The genome was sequenced (Himmelreich et al., 1996) and subsequently reannotated by Peer Bork and colleagues (Dandekar et al., 2000).

5. Obligately intracellular and symbiotic: Tamas et al. (2002) compared the complete genome sequences of two bacteria, *Buchnera aphidicola* (Sg) and *Buchnera aphidicola* (Ap), that are endosymbionts of the aphids *Schizaphis graminum* (Sg) and *Acyrthosiphon pisum* (Ap). Each of these bacteria has a genome size of about 640,000 bp, among the smallest of known bacterial genomes. They have 564 and 545 genes, respectively, of which they share almost all (526). Remarkably, these bacteria diverged about 50 MYA, yet they share complete conservation of genome architecture. There have been no inversions, translocations, duplications, or gene acquisitions in either bacterial genome since their divergence (Tamas et al., 2002). This provides a dramatic example of genomic stasis. Although it is extremely rare for obligate intracellular bacteria to share such genome conservation, it is common for endosymbionts to have relatively small genome sizes. This may reflect the dependence of these bacteria on nutrients derived from the host.

6. Obligately intracellular and parasitic: *Rickettsia prowazekii* is the bacterium that causes epidemic typhus. Its genome is relatively small, consisting of 1.1 Mb (Andersson et al., 1998). Like other *Rickettsia,* it is an α-proteobacterium that infects eukaryotic cells selectively. It is also of interest because it is closely related to the mitochondrial genome. A closely related species, *Rickettsia conorii,* is an obligate intracellular parasite that causes Mediterranean spotted fever in humans. Its genome was sequenced by Ogata et al. (2001). Similar to the *Buchnera aphidicola* subspecies, the genome organization of the two *Rickettsia* parasites is well conserved.

Why are some bacterial genome sizes severely reduced? Intracellular parasites are subject to deleterious mutations and substitutions that cause gene loss, tending toward genome reduction (Andersson and Kurland, 1998). A similar process occurred as a primordial α-proteobacterium evolved into the modern mitochondrion, maintaining only a minuscule mitochondrial genome size (Chapter 12).

Classification of Bacteria Based on Human Disease Relevance

Bacteria and eukaryotes have engaged in an ongoing war for millions of years. Bacteria occupy the nutritive environment of the human body in an effort to reproduce. Typical sites of bacterial colonization include the skin, respiratory tract, digestive tract (mouth, large intestine), urinary tract, and genital system (Eisenstein and Schaechter, 1999). It has been estimated that each human has more bacterial

TABLE 14-5 Vaccine-Preventable Bacterial Diseases

Disease	Species
Anthrax	*Bacillus anthracis*
Diarrheal disease (cholera)	*Vibrio cholerae*
Diphtheria	*Corynebacterium diphtheriae*
Community acquired pneumonia	*Haemophilus influenzae* type B, *Streptococcus pneumoniae*
Lyme disease	*Borrelia burgdorferi*
Meningitis	*Haemophilus influenzae* type B (HIB), *Streptococcus pneumoniae, Neisseria meningitidis*
Pertussis	*Bordetella pertussis*
Tetanus	*Clostridium tetani*
Tuberculosis	*Mycobacterium tuberculosis*
Typhoid	*Salmonella typhi*

Source: Adapted from ▸http://www.cdc.gov/nip/diseases/disease-chart-hcp.htm and ▸http://www.cdc.gov/ncidod/dbmd/diseaseinfo/default.htm.

cells than human cells in the body. In the majority of cases, these bacteria are harmless to humans. However, many bacteria cause infections, often with devastating consequences.

In recent years, the widespread use of antibiotics has led to an increased prevalence of drug resistance among bacteria. It is thus imperative to identify bacterial virulence factors and to develop strategies for vaccination. One approach to this problem is to compare pathogenic and nonpathogenic strains of bacteria (see below). Table 14.5 lists some of the bacterial diseases for which vaccinations are routinely administered. The worldwide disease burden caused by bacteria is enormous.

As of mid-2003, there have been about 150 completed microbial genome-sequencing projects, and another 200 are in progress. At least one representative strain of the major bacteria known to cause human disease are being sequenced.

Worldwide, there are 690,000 new cases of leprosy reported annually; the causative agent is *Mycobacterium leprae*. There are millions of cases of salmonellosis each year, caused by *Salmonella enterica*. A pathogenic strain of *E. coli* (O157:H7) causes haemorrhagic colitis and infects 75,000 individuals in the United States each year. As mentioned above, *M. tuberculosis* infects billions of people and kills millions. You can read about a variety of bacterial diseases at the Centers for Disease Control and Prevention website (▸http://www.cdc.gov/scientific.htm).

An emerging theme in the biology of prokaryotes is that bacterial populations undergo recombination, causing genetic diversification. Joyce et al. (2002) have reviewed this concept in the context of pathogenic bacteria such as *Helicobacter pylori* (a leading cause of gastric ulcers), *Streptococcus pneumoniae,* and *Salmonella enterica.* While eukaryotes achieve genetic diversity through sexual reproduction, prokaryotes also achieve tremendous genetic diversity through other mechanisms such as recombination and lateral gene transfer. Joyce et al. (2002) describe the important role of DNA microarray technology to measure gene expression changes associated with closely related bacterial strains that produce distinct disease phenotypes in humans.

Classification of Bacteria and Archaea Based on Ribosomal RNA Sequences

The main way we know to sample the diversity of microbial life is by molecular phylogeny. Trees have been generated based on multiple sequence alignments of 16S rRNA from various species. Ribosomal RNA has excellent characteristics as a

molecule of choice for phylogeny: It is distributed globally, it is highly conserved yet still exhibits enough variability to reveal informative differences, and it is only rarely transferred between species. An example is shown in Figure 14.1, and we saw a similar tree reconstruction in Figure 12.1.

A major conclusion of the earliest rRNA studies, by Carl Woese and colleagues (Woese and Fox, 1977; Fox et al., 1980), is that bacteria and archaea are distinct groups. The deepest branching phyla are hyperthermophilic microbes, consistent with the hypothesis that the universal ancestor of life existed at hot temperatures (Achenbach-Richter et al., 1987).

A great advance in our appreciation of microbial diversity has come from the realization that the vast majority of bacteria and archaea are noncultivatable (Hugenholtz et al., 1998). It is straightforward to obtain microbes from natural sources and grow some of them in the presence of different kinds of culture medium. For most other microbes, culture conditions are not known. It is still possible to sample uncultivated (or uncultivatable) microbes by extracting nucleic acids directly from naturally occurring habitats (Delong and Pace, 2001). Early studies of archaea described them as extremophiles, but more recent samplings indicate that they are present in nonextreme habitats such as surface waters of the oceans and agricultural soils (DeLong, 1998). Twenty percent of all microbes in the oceans may be archaea (DeLong and Pace, 2001).

Because of sampling bias, four bacterial phyla have been characterized most fully: Proteobacteria, Firmicutes, Actinobacteria, and Bacteroidetes (Hugenholtz, 2002). These major groups account for over 90% of all known bacteria (discussed in Gupta and Griffiths, 2002). However, 35 bacterial and 18 archaeal phylum-level lineages are currently known (Hugenholtz, 2002). Analyses of uncultivated microbes will expand our view of bacterial and archaeal diversity.

The TIGR CMR (see above) lists 12 major divisions of bacteria (Table 14.6) as well as the two major divisions of archaea: crenarchaeota and euryarchaeota (Table 14.7). This classification is similar to that provided at NCBI (Fig. 14.4).

Brochier and Philippe (2002) have contested the view that hyperthermophilic bacteria (such as Aquificales and Thermotaogales) are the most deeply branching. Instead, they suggest that Planctomycetales are positioned at the base of the tree.

Reysenbach and Shock (2002) described a phylogenetic tree of extremophilic microbes based on 16S rRNA sequences. They used a software package designed for rRNA studies, called ARB. You can obtain this software at▸ http://www.arb-home.de/.

Homologous Bacterial Genes Database (HOBACGEN) is a database of proteins from SWISS-PROT and TrEMBL that is organized into families for phylogenetic analysis (Perrière et al., 2000). You can access the database at▸ http://pbil.univlyon1.fr/databases/hobacgen.html.

Classification of Bacteria and Archaea Based on Other Molecular Sequences

In addition to rRNA, many other DNA, RNA, or protein sequences can be used for molecular phylogeny studies. The use of individual proteins (or genes) for such studies often yields tree topologies that conflict with each other and with topologies obtained using rRNA sequences. These discrepancies are usually attributed either to lateral gene transfer (see below), which can confound phylogenetic reconstruction, or to the loss of phylogenetic signals due to saturating levels of substitutions in the gene or protein sequences. A strategy to circumvent this problem is to use combined gene or protein sets. Brown et al. (2001) aligned 23 orthologous proteins conserved across 45 species. Their trees supported thermophiles as the earliest evolved bacteria lineages (Fig. 14.5).

There are many other approaches to bacterial phylogeny. One is to identify conserved insertions and deletions in a large group of proteins. Such "signature sequences" can distinguish bacterial groups and form the basis of a tree (Fig. 14.6) (Gupta and Griffiths, 2002). This tree shows the relative branching order of bacterial species from completed genomes. Eugene Koonin and colleagues (Wolf et al., 2001) used five independent approaches to construct trees for 30 completely

We will study eukaryotes from the perspective of a tree that uses a combined protein data set (Figs. 15.1 and 16.1).

TABLE 14-6 Classification of Bacteria at TIGR Comprehensive Microbial Resource

Bacteria are described as a kingdom, followed by "intermediate ranks." Intermediate rank 3 is not shown

Intermediate Rank 1	Intermediate Rank 2	Genus, Species, and Strain (Examples)	Genome Size (Mb)	GenBank Accession
Actinobacteria	Actinobacteridae	*Mycobacterium tuberculosis* CDC1551	4.4	NC_002755
Aquificae	Aquificales	*Aquifex aeolicus* VF5	1.5	NC_000918
Bacteroidetes	Bacteroides	*Porphyromonas gingivalis* W83	2.3	NA
Chlamydiae	Chlamydiales	*Chlamydia trachomatis* serovar D	1.0	NC_000117
Chlorobi	Chlorobia	*Chlorobium tepidum* TLS	2.1	NC_002932
Cyanobacteria	Chroococcales	*Synechocystis* sp. PCC6803	3.5	NC_000911
	Nostocales	*Nostoc* sp. PCC 7120	6.4	NC_003272
Deinococcus-Thermus	Deinococci	*Deinococcus radiodurans* R1	2.6	NC_001263
Firmicutes	Bacillales	*Bacillus subtilis* 168	4.2	NC_000964
	Clostridia	*Clostridium perfringens* 13	3.0	NC_003366
	Lactobacillales	*Streptococcus pneumoniae* R6	2.0	NC_003098
	Mollicutes	*Mycoplasma genitalium* G-37	0.580	NC_000908
Fusobacteria	Fusobacteria	*Fusobacterium nucleatum* ATCC 25586	2.1	NC_003454
Proteobacteria	Alphaproteobacteria	*Rickettsia prowazekii* Madrid E	1.1	NC_000963
	Betaproteobacteria	*Neisseria meningitidis* MC58	2.2	NC_003112
	Epsilon subdivision	*Helicobacter pylori* J99	1.6	NC_000921
	Gamma subdivision	*Escherichia coli* K12-MG1655	4.6	NC_000913
	Magnetotactic cocci	*Magnetococcus* sp. MC-1	NA	NA
Spirochaetales	Spirochaetaceae	*Borrelia burgdorferi* B31	0.91	NC_001318
Thermotogales	Thermotoga	*Thermotoga maritima* MSB8	1.8	NC_000853

Note: (See ▸http://www.tigr.org/tigr-scripts/CMR2/CMRGenomes.spl). Abbreviation: NA, not available.

sequenced bacterial genomes and 10 sequenced archaeal genomes:

1. They assessed genes that are present or absent in each of these genomes using the COG database (see below). Seventeen invariant genes were identified (all of which encode ribosomal proteins and RNA polymerase subunits). In at least one of the genomes, 4586 additional gene pairs were missing.
2. They assessed the conservation of local gene order (i.e., pairs of adjacent genes) among the genomes.

TABLE 14-7 Classification of Archaea at TIGR Comprehensive Microbial Resource

Archaea are described as a kingdom, followed by "intermediate ranks." Intermediate rank 3 is not shown

Intermediate Rank 1	Intermediate Rank 2	Genus, Species, and Strain (Examples)	Genome Size, (Mb)	GenBank Accession
Crenarchaeota	Thermoprotei	*Aeropyrum pernix* K1	1.6	NC_000854
Euryarchaeota	Archaeoglobi	*Archaeoglobus fulgidus* DSM4304	2.2	NC_000917
	Halobacteria	*Halobacterium* sp. NRC-1	2.0	NC_002607
	Methanobacteria	*Methanobacterium thermoautotrophicum* delta H	1.7	NC_000916
	Methanococci	*Methanococcus jannaschii* DSM2661	1.6	NC_000909
	Methanopyri	*Methanopyrus kandleri* AV19	1.6	NC_003551
	Thermococci	*Pyrococcus abyssi* GE5	1.7	NC_000868
	Thermoplasmata	*Thermoplasma volcanium* GSS1	1.5	NC_002689

Note: (See ▸http://www.tigr.org/tigr-scripts/CMR2/CMRGenomes.spl).

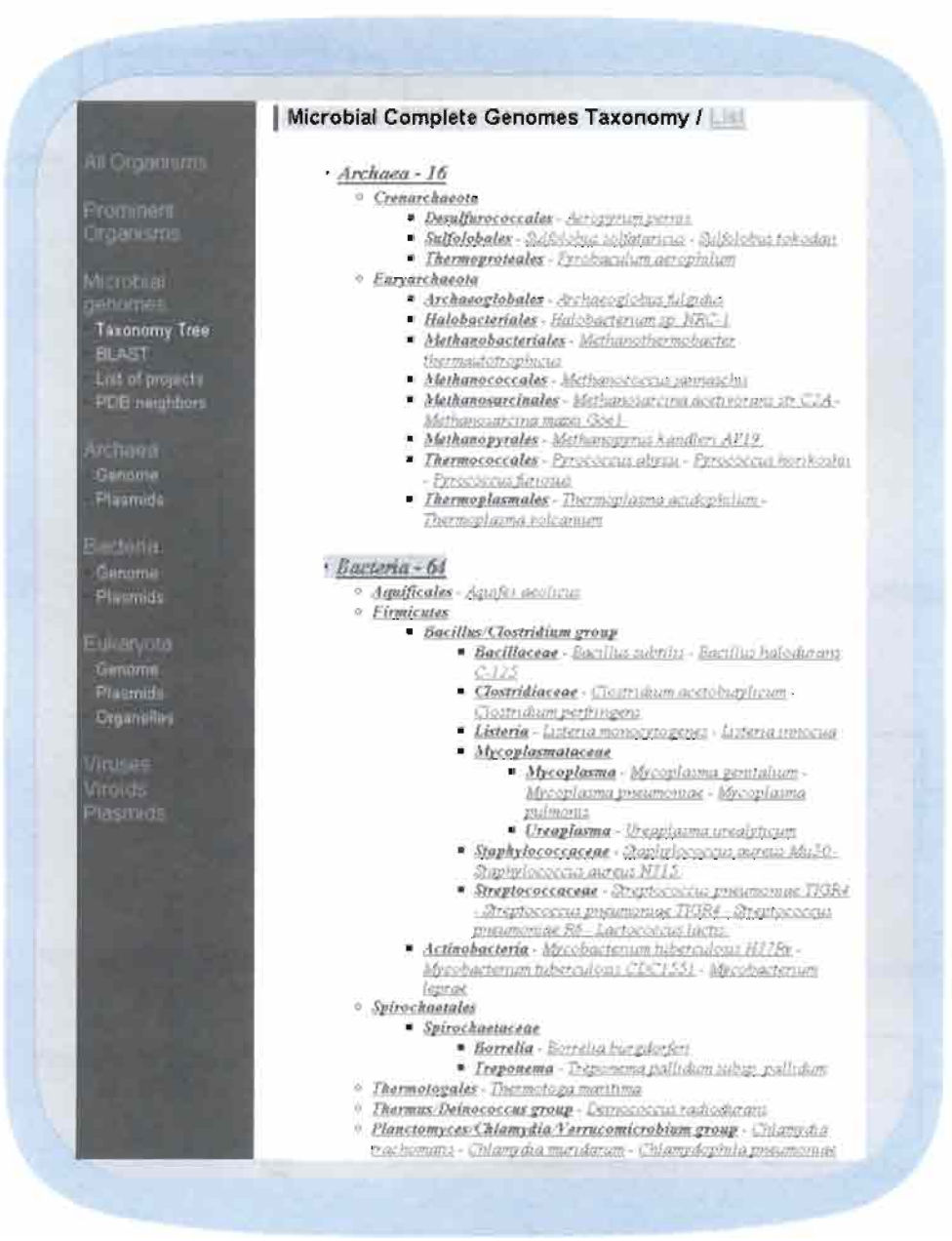


FIGURE 14.4. The NCBI taxonomy tree for microbes includes two major subdivisions of archaea and several subdivisions of bacteria (Proteobacteria and Cyanobacteria divisions are not shown) (http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/new_micr.html).

3. They measured the distribution of percent identity between likely orthologs.
4. They aligned 32 ribosomal proteins into a multiple sequence alignment consisting of 4821 columns (characters) and then generated a tree using the maximum-likelihood approach.
5. They compared multiple trees generated from a series of protein alignments. Wolf et al. (2001) concluded that traditional alignment-based methods were as effective as newer approaches based on genomic data such as local gene order. However, these approaches can yield different kinds of information (e.g., analysis of orthologs can identify genes that have been lost or horizontally transferred between lineages).

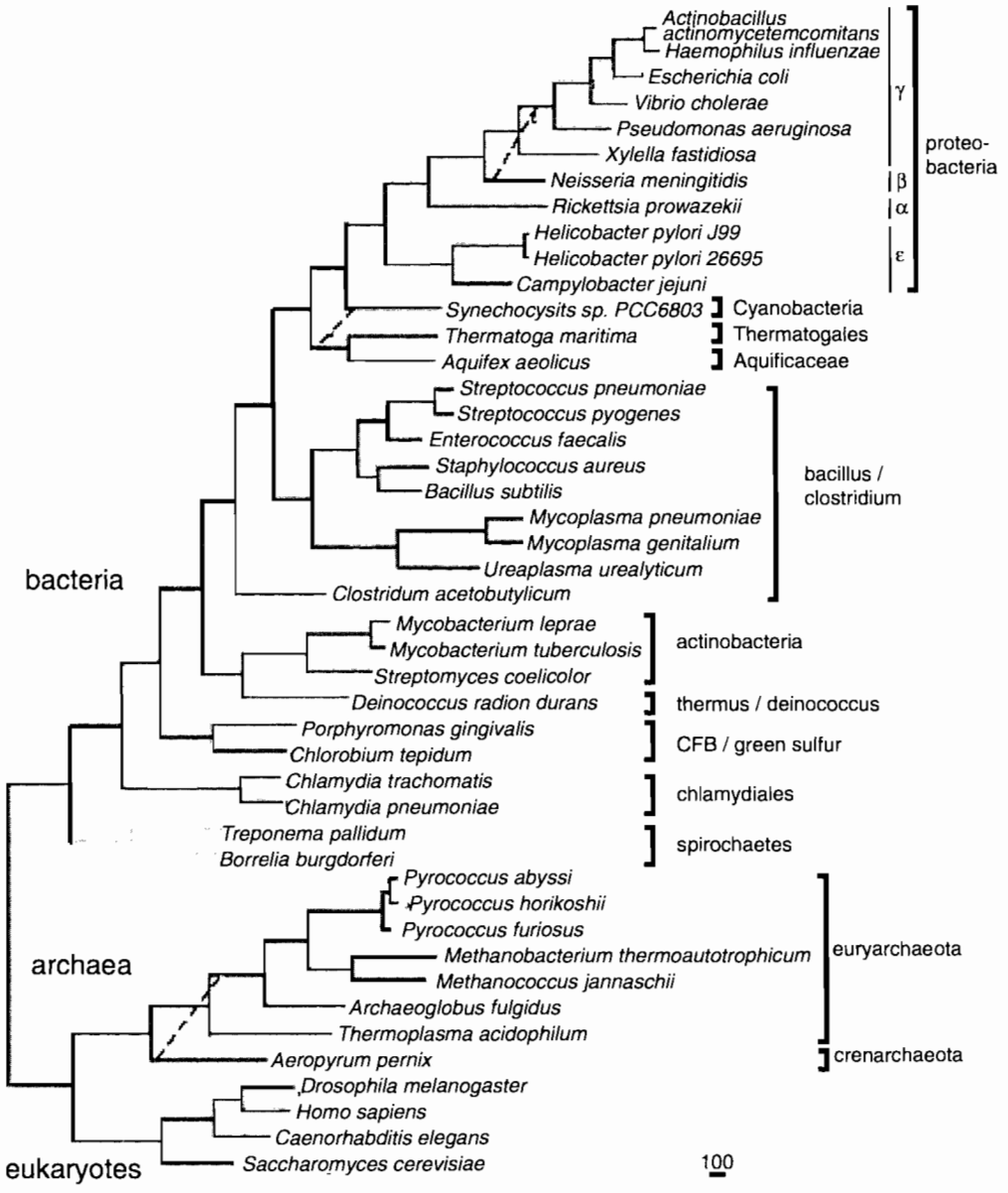


FIGURE 14.5. *An unrooted tree of life adapted from Brown et al. (2001) is based on an alignment of 23 proteins (spanning 6591 amino acid residues). These proteins are conserved across 45 species and include tRNA synthetases, elongation factors, and DNA polymerase III subunit. By combining these proteins, there are many phylogenetically informative sites. The tree consists of three major, monophyletic branches of life as described in Chapter 12. The tree was generated in PAUP by maximum parsimony. Used with permission.*

Analysis of Prokaryotic Genomes

Nucleotide Composition

In the analysis of a completed genome, the nucleotide composition has characteristic properties. The GC content is the mean percentage of guanine and cytosine, and it typically varies from 25 to 75% in eubacteria. Eukaryotes almost always have a larger and more variable genome size than bacteria, but their GC content is very uniform (around 40–45%). Within each species, nucleotide composition tends to be uniform.

We showed the range of GC content in Figure 12.16.

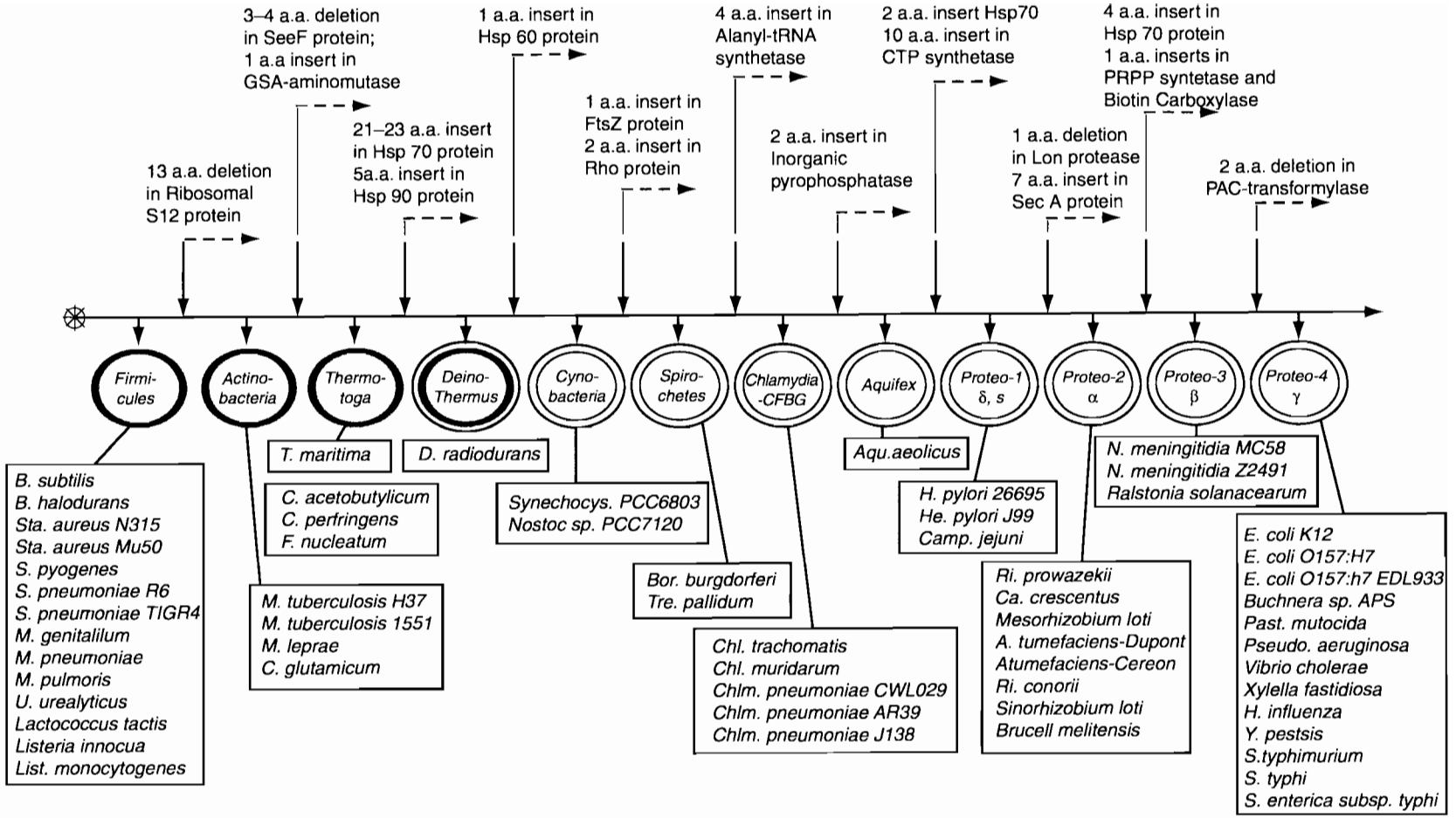


FIGURE 14.6. *Phylogenetic relationships of bacterial species based on signature sequences in different proteins. The solid arrows above the line indicate the evolutionary point at which each insertion or deletion is proposed to have occurred. Thus, all bacterial groups to the right of each arrow are predicted to share that signature insertion/deletion while all groups to the left lack those changes. From Gupta and Griffiths (2002). Used with permission.*

TABLE 14-8 Programs for Gene Finding in Prokaryotic Genomes

Program	Description	URL
Critica	From Jonathan Badger, University of Waterloo	▸http://www.math.uwaterloo.ca/~jhbadger/critica/critica105/critica.html
ECOPARSE	A mail server, from Anders Krogh	▸http://www.cbs.dtu.dk/~krogh/EcoParse.info
FrameD	Locates genes and frameshifts; optimized for GC-rich genomes	▸http://genopole.toulouse.inra.fr/
GeneMark.hmm	Uses hidden Markov models	▸http://intron.biology.gatech.edu/GeneMark
GeneMarkS	Improved accuracy of predicting gene start sites	▸http://dixie.biology.gatech.edu/GeneMark/genemarks.cgi
GLIMMER	At TIGR	▸http://www.tigr.org/softlab/glimmer/glimmer.html
Orpheus	Gene prediction in complete bacterial genomes and large genomic fragments	▸http://pedant.gsf.de/orpheus/

Two-dimensional bacterial genomic display (2DBGD) is an electrophoresis technique analogous to two-dimensional protein gels (Chapter 8) in which bacterial genomic DNA is fragmented and separated first by size and then by sequence composition. Malloff et al. (2002) used 2DBGD to compare the genomes of three respiratory pathogens with very high GC content: *Bordetella pertussis,* the cause of whooping cough (68% GC); *M. tuberculosis* (66% GC); and *Mycobacterium avium* (66% GC). This technique can be used to detect insertions and deletions in prokaryotic strains.

You can access online programs to display GC content at L'Atelier BioInformatique de Marseille (▸http://www.up.univ-mrs.fr/~wabim/d_abim/riche-adn.html) or with the Emboss program GEECEE (▸http://bioweb.pasteur.fr/seqanal/interfaces/geecee.html). Many other programs show GC content, such as GLIMMER (see below).

Fractionated genomic DNA can be separated into fractions of different buoyant densities by centrifugation. Homogenous stretches of DNA, called isochores, band at characteristic densities and reflect regions high in either AT content or GC content. Genes tend to be localized to heavy isochores associated with a high GC content. We will discuss isochores in our analysis of the human genome (Chapter 17).

Finding Genes

Bacteria and archaea are characterized by a high gene density (about one gene per kilobase), absence of introns, and very little repetitive DNA. Thus the problem of finding genes is relatively simple in comparison to searching eukaryotic DNA (Chapters 15–17). Several programs are available for microbial gene identification (Table 14.8) (reviewed in Baytaluk et al., 2002).

There are four main features of genomic DNA that are useful for gene recognition (Baytaluk et al., 2002). These features apply to both bacterial and eukaryotic gene finding:

1. *Open reading frame (ORF) length.* An ORF is not necessarily a gene; for example, many short ORFs are not part of authentic genes (discussed further below). An ORF is defined by a start codon (i.e., ATG encoding a methionine) and a stop codon (TAA, TAG, TGA). However, in bacteria, alternative start codons may be employed such as GTG or TTG, and there are rarely used alternative stop codons.

2. *Presence of a consensus sequence for ribosome binding in the immediate vicinity of the start codon.* In some cases, it is possible to identify two in-frame ATG codons, either of which could represent the start codon. Identifying a ribosome binding site can be an important indicator of which is the likely start site. In bacteria, the ribosome binding site is called a Shine–Dalgarno sequence. It is a purine-rich stretch of nucleotides that is complementary to the 3′ end of 16S rRNA, extending from the −20 position (i.e., 5′ to the initiation codon) to the +13 position (i.e., 13 nucleotides downstream in the 3′ direction). Samuel Karlin and colleagues (Ma et al., 2002) studied 30 prokaryotic genomes and correlated the features of the Shine–Dalgarno sequence with expression levels of genes based on codon usage bias (see below), type of codon, functional gene class, and type of start codon. They have shown a positive correlation between the presence of a strong Shine–Dalgarno sequence and high levels of gene expression.

3. *Presence of a pattern of codon usage that is consistent with genes.* Hidden Markov models (Chapter 10) have been particularly useful in defining the coding potential of putative protein-coding DNA sequences.

4. *Homology of the putative gene to other, known genes.* Genomic DNA sequences, including putative genes, can be searched against protein databases using blastx (see Chapter 5). This approach is especially helpful in finding genes in eukaryotic organisms. For example, exons can be matched to expressed sequence tags (Chapter 16).

The first three of these features are studied using intrinsic approaches to gene finding. They are called intrinsic because the features do not necessarily depend on comparisons to gene sequences from other organisms. The fourth feature, relationship to other genes, is called an extrinsic approach. Some prokaryotic gene-finding programs, such as Critica (Table 14.8), combine both intrinsic and extrinsic approaches.

Intrinsic approaches are also sometimes called ab initio approaches.

The GLIMMER system is one of the premier gene-finding algorithms. It identifies about 97–98% of all genes in a bacterial genome (Delcher et al., 1999a). The algorithm uses interpolated Markov models (IMMs). A Markov chain can describe the probability distribution for each nucleotide in a genomic DNA sequence (see Chapter 10). This probability can depend on the preceding k variables (nucleotides) in the sequence. A fixed-order Markov chain would describe the k-base context for each nucleotide position; for example, a fixed fifth-order Markov chain model describes $4^5 = 1024$ probability distributions, one for each possible 5-mer. The k-mers are used as a training set to teach the algorithm the rules for which probability distributions are most likely to be relevant to this particular genomic sequence. Larger values for k are more informative. IMMs are a specialization of Markov models in which rare k-mers tend to be ignored, and more common k-mers are weighted more heavily. The algorithm generates a score for each potential coding region.

GLIMMER was written by Owen White and colleagues at TIGR. GLIMMER is an acronym for Gene Locator and Interpolated Markov Modeler.

To use GLIMMER, it is necessary to run the software on a UNIX operating system. First, enter a data set of genomic DNA from the organism of interest (e.g., *E. coli*) in order to train the algorithm. The command to do this is shown in Figure 14.7*a*, and the output for the analysis of 76,000 nucleotides of *E. coli* genomic DNA is shown in Figure 14.7*b*. This shows the GC content of the DNA fragment (51.5% in this case), the parameters (e.g., the minimum gene length is set to 90 nucleotides), and a list of the predicted genes, including orientation (on the forward or reverse strand), length, and score. The GLIMMER output also has a summary of the predicted genes, including notations on possible overlaps (Fig. 14.7*c*).

There are several pitfalls associated with prokaryotic gene prediction:

- There may be multiple genes that are encoded by one genomic DNA segment, in an alternate reading frame on the same strand or opposite strand.
- It is difficult to assess whether a short ORF is genuinely transcribed. According to Skovgaard et al. (2001), there are far too many short genes annotated in many genomes. For *E. coli,* they suggest that there are 3800 true protein-coding genes rather than the 4300 genes that have been annotated. Since stop codons (TAA, TAG, TGA) are AT rich, genomes that are GC rich tend to have fewer stop codons and more predicted long ORFs. For all predicted proteins in a genome, the proportion of hypothetical proteins (defined as predicted proteins for which there is no experimental evidence that they are expressed) rises greatly as sequence length is smaller.

(*a*)

```
$ /usr/local/glimmer2.02/build-icm < ecoli_first100.txt > trainecoli
$ /usr/local/glimmer2.02/glimmer2 ecoli76k.fasta colitrain
```

(*b*)

```
GC Proportion = 51.5%
Minimum gene length = 90
Minimum overlap length = 30
Minimum overlap percent = 10.0%
Use independent scores = True
Ignore independent score on orfs longer than 765
Use first start codon = True
```

ID#	Fr	Orf Start	Gene Start	End	Lengths Orf	Lengths Gene	Gene Score	-- Frame Scores - F1	F2	F3	R1	R2	R3	Indep Score	
	F2	35	44	178	144	135	0	0	0	0	_	0	_	99	0
	F1	226	247	402	177	156	0	0	_	_	_	0	_	99	0
	F3	273	420	563	291	144	0	_	_	0	_	_	0	99	0
	R2	740	713	609	132	105	0	_	6	_	_	0	2	89	0
	F2	515	548	916	402	369	0	_	0	_	_	_	0	99	6
	R1	1149	1143	1036	114	108	0	_	_	0	0	_	0	99	0
	F3	888	936	1265	378	330	0	_	_	0	_	_	0	99	0
	F1	1162	1210	1347	186	138	0	0	_	_	_	_	0	99	0
	F3	1365	1377	1592	228	216	0	_	_	0	_	_	0	99	0
	F3	1707	1710	1823	117	114	0	_	_	0	_	_	0	99	0
1	R3	1951	1909	380	1572	1530	99	_	_	_	_	_	99	0	101
	F3	1857	1872	1994	138	123	0	_	_	0	_	_	_	99	0
	R1	2124	2121	2029	96	93	0	_	_	_	0	_	_	99	0
	F3	2043	2043	2273	231	231	0	_	_	0	_	_	0	99	0
	F1	2098	2140	2319	222	180	0	0	0	_	_	_	0	99	0
	R1	2604	2589	2182	423	408	0	_	0	_	0	_	0	99	0
	F3	2313	2349	2645	333	297	0	_	0	0	_	_	0	99	0
	F2	2057	2183	2692	636	510	0	_	0	_	_	_	0	99	0
2	R1	2844	2835	2692	153	144	90	_	_	_	90	_	1	7	90
	F1	2809	2815	2946	138	132	7	7	_	0	_	_	4	87	7
	F3	2769	2835	2987	219	153	0	_	_	0	_	_	5	94	0
3	R3	3034	2971	2039	996	933	99	_	_	_	_	_	99	0	109

(*c*)

```
     48    66485    65460  [-3 L=1026]
     49    69967    68642  [-2 L=1326]
     50    71402    70614  [-3 L= 789]
     51    73146    71341  [-1 L=1806]
     52    73426    74445  [+1 L=1020]
     53    74625    76166  [+3 L=1542]
     54    80198    76878  [-3 L=3321]
     55    80118    81140  [+3 L=1023]  [LowScoreBy #56 L=918 S=2]
     56    81188    80223  [-3 L= 966]  [OlapWith #55 L=918 S=97]
     57    82370    81288  [-3 L=1083]
     58    84895    83759  [-2 L=1137]  [DelayedBy #59 L=108]
     59    84873    86438  [+3 L=1566]
```

FIGURE 14.7. *The GLIMMER program (from TIGR; http://www.tigr.org) is useful to find genes in bacterial DNA. The program is run on a UNIX operating system.* (a) *A data set of genomic DNA must first be trained to generate Markov models, and then the program is run.* (b) *The output includes a list of identifiers with ORF data on the forward (F1, F2, F3) and reverse (R1, R2, R3) strands and scores for the likelihood that a gene has been identified.* (c) *The output also includes a list of several genes. The DNA used was from* E. coli *K12 (accession number U14003).*

An operon is a cluster of contiguous genes, transcribed from one promoter, that gives rise to a polycistronic mRNA. The predicted gene pairs from this study, encompassing 73 bacterial and archaeal genomes, are available on the web at http://www.tigr.org/tigr-scripts/operons/operons.cgi.

- Frameshifts can occur, in which the genomic DNA is predicted to encode a gene with a stop codon in one frame but a continuing sequence in another frame on the same strand. A frameshift could be present because of a sequencing error or because of a mutation that leads to the formation of a pseudogene (a nonfunctional gene).
- Some genes are part of operons that often have related functional roles in prokaryotes. Operons have promoter and terminator sequence motifs, but

these are not well characterized. Steven Salzberg and colleagues (Ermolaeva et al., 2001) analyzed 7600 pairs of genes in 34 bacterial and archaeal genomes that are likely to belong to the same operon.

- Lateral gene transfer, also called horizontal gene transfer, commonly occurs in bacteria and archaea. We will discuss this next.

Lateral Gene Transfer

Lateral, or horizontal, gene transfer (LGT) is the phenomenon in which a genome acquires a gene from another organism directly, rather than by descent (Eisen, 2000; Koonin et al., 2001). It is possible to identify a gene that is most closely related to orthologs in distantly related organisms. The simplest explanation for how a species acquired such a gene is through lateral gene transfer. This mechanism represents a major force in genome evolution. The gene transfer is unidirectional, rather than involving a reciprocal exchange of DNA.

Lateral gene transfer is a significant phenomenon for several reasons:

1. This mechanism vastly differs from the normal mode of inheritance in which genes are transmitted from parent to offspring. Thus, lateral gene transfer represents a major shift in our conception of evolution.
2. This mechanism is very common in prokaryotes, and many examples have been described in eukaryotes as well. It has been observed within and between each of the three main branches of life but is particularly prevalent in bacteria.
3. Lateral gene transfer can greatly confound phylogenetic studies. If a DNA, RNA, or protein is selected for phylogenetic analysis that has undergone lateral gene transfer, then the tree will not accurately represent the natural history of the species under consideration. An extreme interpretation of lateral gene transfer is that if it is common enough, then it is impossible in principle to derive a single true tree of life.

Lateral gene transfer occurs as a multistep process (Fig. 14.8) (Eisen, 2000). A gene that involves in one lineage (by the traditional Darwinian process of vertical descent) may transfer to the lineage of a second species. This DNA transfer could be mediated by a viral vector or by a mechanism such as homologous recombination. Once a new gene is incorporated into the genome of individuals with a population (e.g., species 3 in Fig. 14.8), positive selection maintains its presence within those individuals. A transferred gene presumably must confer benefits to the new species in order to be maintained, propagated, and spread throughout the population of the new species. Finally, the new gene adapts to its new lineage, a process called "amelioration" (Eisen, 2000) (Fig. 14.8, arrow 6).

Carl Woese (2002) has suggested that in early evolution lateral gene transfer predominated to such an extent that primitive cellular evolution was a communal process, followed only later by vertical (Darwinian) evolution.

As a specific example of a gene that has undergone lateral gene transfer, look at a human protein, thymidine phosphorylase (P19971). This protein is encoded by one of several dozen genes that Salzberg et al. (2001) identified as candidates for lateral gene transfer from bacteria. A blastp search with this protein results in many dozens of database matches with low expect values (high scores). As shown in the taxonomy report, the best match is to itself (Fig. 14.9, arrow 1) and to mouse and rat orthologs (arrows 2 and 3). All other database matches are to proteins from prokaryotes such as *E. coli* and *Bacillus anthracis*. Thus this gene may have entered the human lineage by lateral gene transfer. Upon the publication of the draft sequence of the

You can see a list of several dozen human genes that may have been incorporated into the human genome at the supplemental website to Salzberg et al. (2001) (http://www.sciencemag.org/cgi/data/1061036/DC1/1).

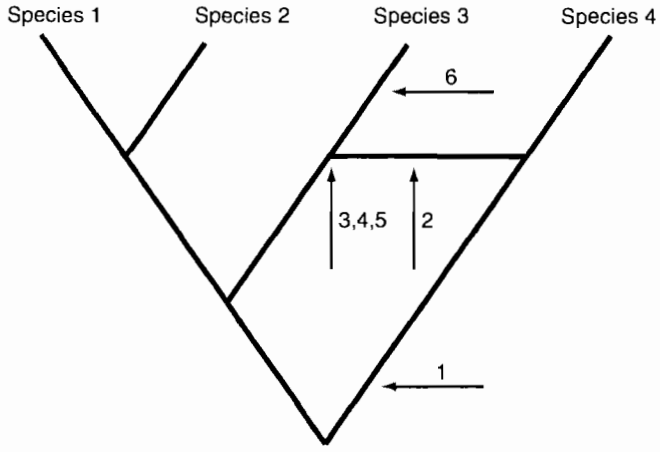


FIGURE 14.8. *Lateral gene transfer occurs in stages. In this hypothetical scenario, four species evolved from a common ancestor. Genes in each species descend in a horizontal fashion over time (arrow 1). At some point in time, a gene transfers horizontally from the lineage of species 4 to the lineage of species 3 (arrow 2). Transferred genes must then be fixed in some individual genomes (arrow 3), maintained under strong positive selection (arrow 4), and spread through the population of species 3 (arrow 5). The laterally transferred gene then evolves as an integral part of the new genome (arrow 6). This gene may be distinguished from other genes in species 3 by having a nucleotide composition or codon usage profile that is characteristic of species 4. This figure is adapted from Eisen (2000). Used with permission.*

Organism Report

Homo sapiens (man) [mammals] taxid 9606 — 1
sp|P19971|TYPH_HUMAN Thymidine phosphorylase precursor (Td... 941 0.0
gb|AAA60043.1| endothelial cell growth factor >gi|6862560|... 938 0.0
gb|AAH18160.1|AAH18160 endothelial cell growth factor 1 (p... 937 0.0

Mus musculus (mouse) [mammals] taxid 10090 — 2
ref|NP_612175.1| endothelial cell growth factor 1; thymidi... 732 0.0

Rattus norvegicus (brown rat, ...) [mammals] taxid 10116 — 3
ref|XP_235562.1| similar to thymidine phosphorylase [Mus m... 695 0.0

Deinococcus radiodurans [eubacteria] taxid 1299
pir||A75520 pyrimidine-nucleoside phosphorylase - Deinococ... 322 1e-86
ref|NP_295490.1| anthranilate synthase component II [Deino... 62 2e-08
sp|Q9RTJ5|TRPD_DEIRA Anthranilate phosphoribosyltransferase 60 4e-08

Mycobacterium tuberculosis CDC1551 [high GC Gram+] taxid 83331
ref|NP_337943.1| thymidine phosphorylase [Mycobacterium tu... 306 4e-82

Geobacillus stearothermophilus [eubacteria] taxid 1422
pdb|1BRW|A Chain A, The Crystal Structure Of Pyrimidine Nu... 299 7e-80
pir||JT0875 pyrimidine-nucleoside phosphorylase (EC 2.4.2.... 275 2e-72
gb|AAD33792.1|AF139534_3 anthranilate phophoribosyl transf... 51 4e-05

Sinorhizobium meliloti [a-proteobacteria] taxid 382
ref|NP_384236.1| PROBABLE THYMIDINE PHOSPHORYLASE TRANSMEM... 298 2e-79
sp|Q92PS0|TRPD_RHIME Anthranilate phosphoribosyltransferas... 51 3e-05

Thermoanaerobacter tengcongensis [eubacteria] taxid 119072
ref|NP_622143.1| Thymidine phosphorylase [Thermoanaerobact... 296 5e-79
sp|Q8R9M6|TRPD_THETN Anthranilate phosphoribosyltransferas... 52 2e-05

Streptomyces coelicolor A3(2) [high GC Gram+] taxid 100226
ref|NP_629043.1| thymidine phosphorylase [Streptomyces coe... 295 1e-78

Salmonella enterica subsp. enterica serovar Typhi [enterobacteria] taxid 90370
pir||AD1073 thymidine phosphorylase (EC 2.4.2.4) [imported... 286 4e-76
pir||AE0653 anthranilate synthase component II, anthranila... 51 4e-05

Escherichia coli CFT073 [enterobacteria] taxid 199310
ref|NP_757312.1| Thymidine phosphorylase [Escherichia coli... 284 2e-75
gb|AAN80194.1|AE016760_53 Anthranilate synthase component ... 52 2e-05

Escherichia coli O157:H7 EDL933 [enterobacteria] taxid 155864
ref|NP_290996.1| thymidine phosphorylase [Escherichia coli... 284 2e-75

FIGURE 14.9. *Taxonomy report of a tblastn search using human P19971 (thymidine phosphorylase) as a query. This gene was considered a candidate for lateral gene transfer from bacteria to humans. The database matches include humans (arrow 1), mouse (arrow 2), rat (arrow 3), and a variety of bacteria. No other (lower) eukaryotes have database matches for this query.*

human genome in 2001 (International Human Genome Sequencing Consortium, 2001; Ponting, 2001), it was suggested that several hundred human genes arrived by lateral gene transfer, although subsequent analyses suggested that the actual number is far smaller (Stanhope et al., 2001; Salzberg et al., 2001).

How is lateral gene transfer identified? The main criterion is that a gene has an unusual nucleotide composition, codon usage, phylogenetic position, or other feature that distinguishes it from most other genes in a genome. There are three principal methods by which lateral gene transfer may be inferred:

1. Phylogenetic trees of different genes may be compared. This is the favored approach (Eisen, 2000).
2. Patterns of best matches for each gene in a genome may be used.
3. The distribution pattern of genes across species can be assessed to search for genes that have undergone lateral gene transfer. For example, if a gene is present in crenarcheota and plants but not in other archaea, bacteria, or eukaryotes, this may be taken as evidence favoring a lateral gene transfer mechanism. This approach may lead to artifactual results (false positives) if gene loss or rapid mutation has occurred but not lateral gene transfer.

Let us return to the example of thymidine phosphorylase. There are several reasons for caution in assigning a mechanism of lateral gene transfer:

- If that gene were present in an insect such as *Drosophila* or a plant, then the argument in favor of lateral gene transfer would be considerably weakened. Is thymidine phosphorylase present in chimpanzee? We will not know the answer until that genome is fully sequenced (or unless we make a directed effort to clone that gene from chimpanzee DNA). A major concern in positing lateral gene transfer is that the candidate gene might be present throughout the tree of life, but we might have insufficient sequence data to find it in other species. A tblastn search of thymidine phosphorylase protein against the expressed sequence tag database reveals that there are indeed many eukaryotes that possess this gene (Table 14.9). Thus in this case the gene probably did not undergo lateral gene transfer. Even among bacterial and archaeal genomes, we are only beginning to sample the diversity of complete genomes, and over time it will be progressively easier to assess evolutionary relationships.
- It is also possible that the gene in question has undergone rapid mutation, such that the phylogenetic signal is lost.

TABLE 14-9 Results of tblastn Search of Expressed Sequence Tag Database Using Human Thymidine Phosphorylase as Query (December 2002)

Organism	Common Name	Accession	Expect Value
Salmo salar	Atlantic salmon	BG935885	2×10^{-72}
Oncorhynchus mykiss	Rainbow trout	CA348293	3×10^{-68}
Gallus gallus	Chicken	BM491333	1×10^{-56}
Tetrahymena thermophila	Tetrahymena	BM400607	8×10^{-47}
Drosophila melanogaster	Fruit fly	BI591052	3×10^{-27}
Rattus norvegicus	Rat	AW914201	2×10^{-22}

- Another scenario is that a gene once existed in other eukaryotic lineages but was subsequently lost.

Annotation and Comparison

As prokaryotic genomes are sequenced, they are annotated (see Chapter 12). This process is far more straightforward in bacteria and archaea than in eukaryotes. A large collection of tools is available at TIGR CMR, at NCBI, and at EBI. We begin with NCBI. The Clusters of Orthologous Genes (COG) database organizes information collected from dozens of prokaryotic genomes as well as the yeast *Saccharomyces cerevisiae* (Chapter 15) (Tatusov et al., 1997, 2001). The goal of the COG project is to provide a phylogenetic classification of prokaryotic proteins. The approach is to classify the relationships of proteins in groups based on "best-hit" BLAST search results.

The COG URL is ▸ http://www.ncbi.nlm.nih.gov/COG/. There are currently over 100,000 proteins organized into about 75,000 clusters of orthologous genes.

The COG main page is shown in Figure 14.10. The COGnitor link allows you to enter a sequence of interest to perform a search against the COG database. A text query is available; by entering "lipocalin," you link to a group of prokaryotic lipocalins (classified as COG3040). This shows the relationship of a group of bacterial lipocalins in various species (Fig. 14.11*a*) as well as a multiple sequence

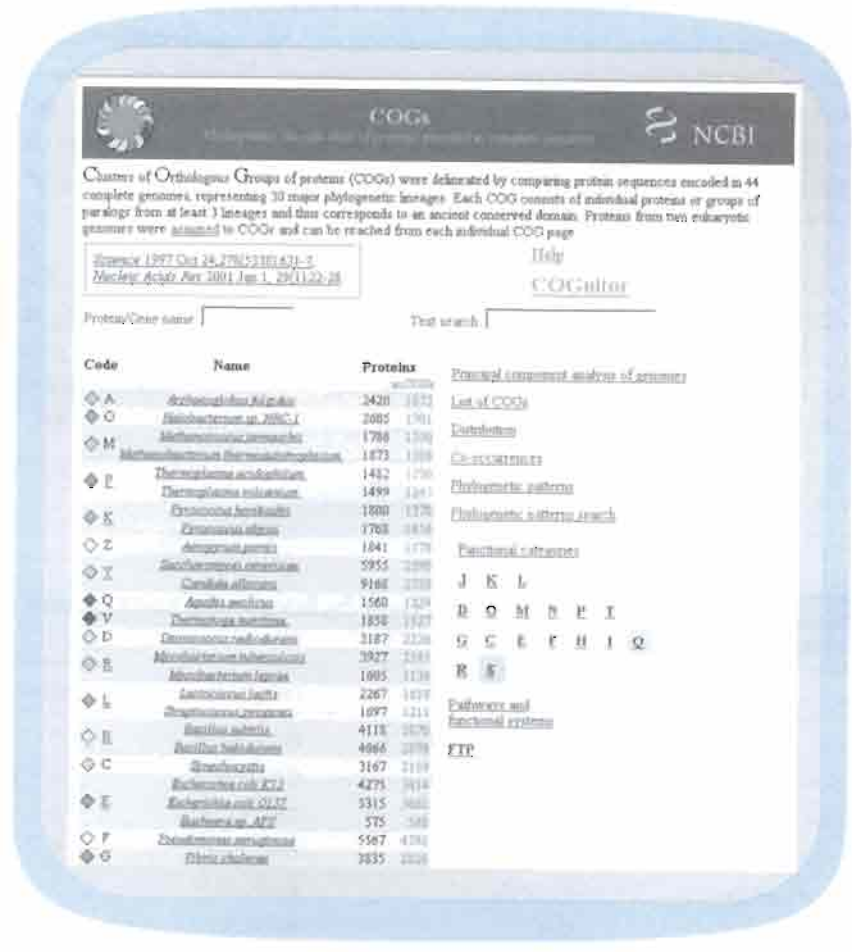
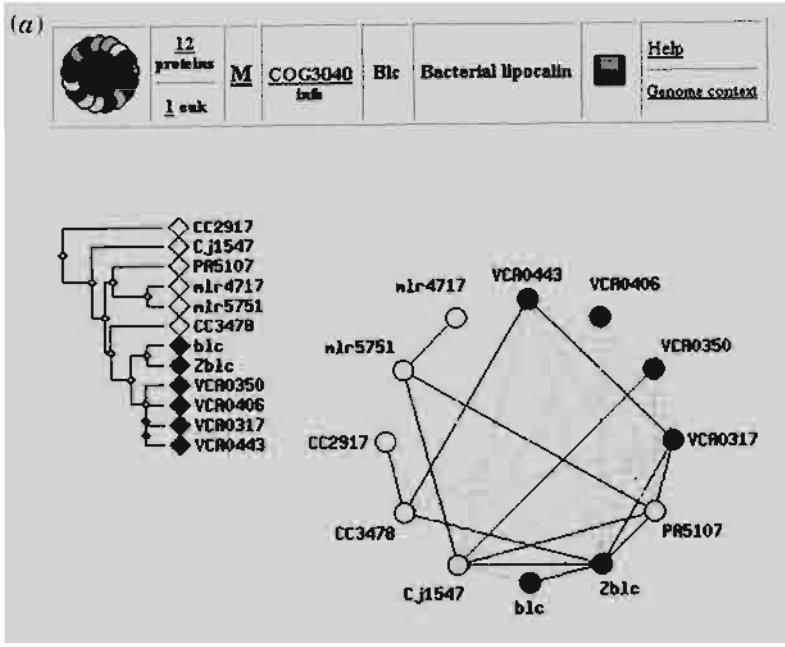


FIGURE 14.10. The COG page at NCBI provides analyses of functionally related genes and proteins from completely sequenced genomes (▸ http://www.ncbi.nlm.nih.gov/COG/).



(b)

```
CLUSTAL V (1.74) multiple sequence alignment

blc             MRLLPLVLAAATAAF--LVVACSSP--TPPRGVTVVNNFDAKRYLGTWYEIARFDHRFERG
Zblc            MRLLPLVLAAATAAF--LVVACSSP--TPPRGVTVVNNFDAKRYLGTWYEIARFDHRFERG
VCA0317         MRAIFLILCA---V--LLNGCLG----HPESVKPVSDFELNNYLGKWYEVARLDHSFERG
VCA0443         MRAIFLILCS---V--LLNGCLG----HPESVKPVSDFELNNYLGKWYEVARLDHSFERG
VCA0406         MRAIFLILCS---V--LLNGCLG----HPESVKPVPDFELNNYLGKWYEVARLDHSFERG
VCA0350         MRAIFLILCS---V--LLNGCLG----HPESVKFVSDFELNNYLGKWYEVARLDHSFEKG
nlr4717         MFGGWIDRKE---------NFNTD----TAAVTAITS-LNLSRYLGKWYELCRLPIKYEDE
nlr5751         MFRGWIDGKE---------NFNTD----TAAVTAITS-LNLSQYLGKWYELCRLPIKYEDE
PA5107          MIGLPIRVLCTLWLGLLLVGCAG----SPKNPPATEQVDLQRYQGTWYELARLPNFFQRN
CC3478          MRKSLLALPVILSLTTTLGACASGPVGWFNSPQPAKAVELDRYLGKWYEVARYDHRFQKG
CC2917          MLNTAIALFA---------LAQAAAV---QPQTPLIDP---ARHYSGVVSEVGRTPERITKG
Cj1547          ------------------------------MLELIGELSLQNYNGEVLEIARKPAFFQKS
                                               .* * * *: *
```

FIGURE 14.11. (a) *Selecting the COG link for "bacterial lipocalins" results in a description of 12 bacterial proteins and 1 eukaryotic protein (from* Drosophila*). The relation of the 12 proteins is shown in two kinds of clustering trees.* (b) *By clicking on the link for the 12 proteins, a multiple sequence alignment generated by ClustalW is provided.*

alignment of these proteins (Fig. 14.11*b*). The notation "M" (Fig. 14.11*a*) indicates the functional classification of each cluster of orthologous genes (M corresponds to outer membrane proteins, such as the bacterial lipocalins).

The COG functional categories can be accessed from the main page (Fig. 14.10), and they are shown in Figure 14.12. This resembles a gene ontology system (Chapter 8); here, the main functional categories are information storage and processing, cellular processes, and metabolism.

We presented the COG functional categories in Table 8.9. (p. 247)

The distribution of clusters of orthologous genes as a function of the number of species is a useful way to identify groups of related proteins that occur very rarely or very frequently. For example, there are 84 clusters of orthologous genes that are found in 26 different species (Fig. 14.13, bottom). These well-conserved protein families include tRNA synthetases, ribosomal proteins, and other enzymes such as signal recognition particle GTPase and S-adenosyl methionine-dependent methyltransferases. Each of these is thus a good candidate for phylogenetic studies. At the other extreme, there are 493 clusters of orthologous genes that are found in just three of the organisms represented in the COG database (Fig. 14.13, top). These entries include many proteins that are annotated as "uncharacterized" or "predicted" (e.g., "uncharacterized membrane protein"). If you are studying a particular

Code	COGs	Domains	Description	Pathways and functional systems
Information storage and processing				
J	217	6449	Translation, ribosomal structure and biogenesis	4
K	132	5438	Transcription	3
L	184	5337	DNA replication, recombination and repair	2
Cellular processes				
D	32	842	Cell division and chromosome partitioning	-
O	110	3165	Posttranslational modification, protein turnover, chaperones	-
M	155	4079	Cell envelope biogenesis, outer membrane	1
N	133	3110	Cell motility and secretion	2
P	160	5112	Inorganic ion transport and metabolism	1
T	97	3627	Signal transduction mechanisms	-
Metabolism				
C	224	5594	Energy production and conversion	7
G	171	5262	Carbohydrate transport and metabolism	4
E	233	8383	Amino acid transport and metabolism	10
F	85	2364	Nucleotide transport and metabolism	5
H	154	4057	Coenzyme metabolism	11
I	75	2609	Lipid metabolism	2
Q	62	2754	Secondary metabolites biosynthesis, transport and catabolism	-
Poorly characterized				
R	449	11948	General function prediction only	-
S	750	6416	Function unknown	-

FIGURE 14.12. A primary feature of the COG database is a functional annotation of proteins into 18 categories. This annotation is important in determining the biochemical pathways that operate in the cells of organisms with completely sequenced genomes.

prokaryote, it could be of interest to study these proteins because they may be relatively unique to that organism.

Comparison of Prokaryotic Genomes

COG

One of the most important lessons of whole-genome sequencing is that comparative analyses greatly enhance our understanding of genomes. It can be useful to compare genomes whether they are closely or distantly related organisms. Some of the species that have had the genomes of closely related strains completely sequenced are indicated in Table 14.10. It will be significant to compare such genomes for several reasons:

- We may be able to discover why some strains are pathogenic.
- Eventually, we may be able to predict clinical outcome of infections based on the genotype of the pathogen.
- We may develop strategies for vaccine development.

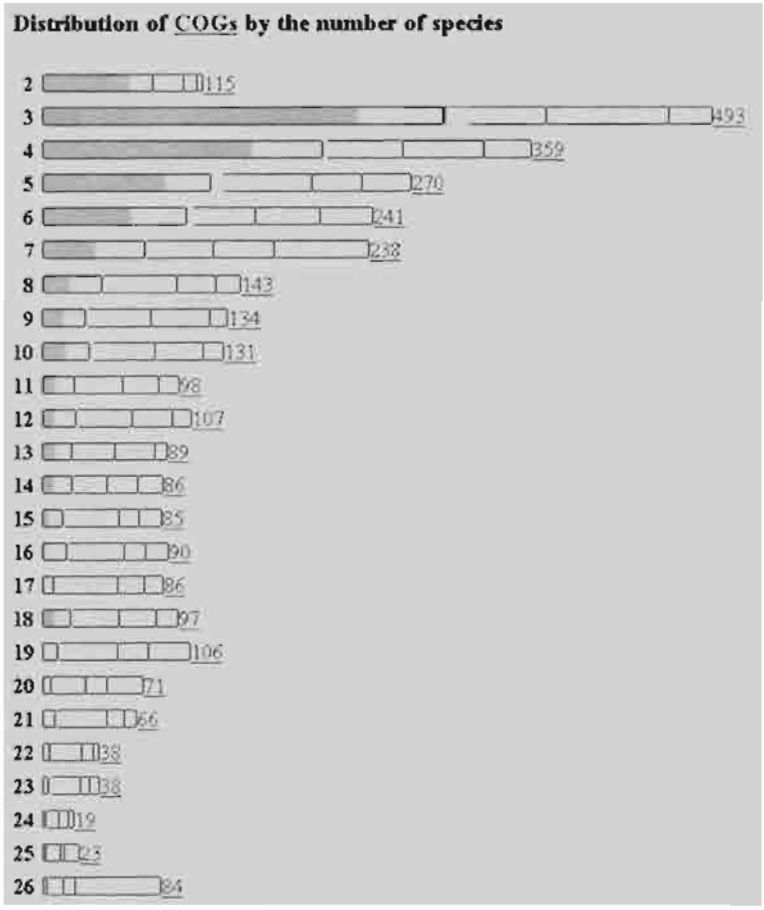


FIGURE 14.13. *The COG distribution shows that the majority of the clusters are present in 3–7 species. Eighty-four clusters of orthologous genes are present in all 26 species represented in the COG database; these 84 clusters include tRNA synthetases, ribosomal proteins, kinases, and other proteins that are extremely well conserved among most life forms on the planet.*

Continuing our discussion of COG, this database features a phylogenetic pattern search (see Fig. 14.10). This allows you to compare species to each other. From the search page (Fig. 14.14) select *Chlamydia trachomatis* and *Chlamydia pneumoniae*. Chlamydiae are obligate intracellular bacteria that are phylogenetically distinct from other bacterial divisions. *Chlamydia pneumoniae* infects humans, causing pneumonia and bronchitis. *Chlamydia trachomatis* causes trachoma (an ocular disease that leads to blindness) and sexually transmitted diseases. Why do these closely related bacteria affect different body regions and cause such distinct pathologies? Their genomes have been sequenced and compared (Stephens et al., 1998; Kalman et al., 1999; Read et al., 2000).

In the United States, 10% of all pneumonia cases and 5% of bronchitis cases are attributed to *C. pneumoniae*.

We can compare the functionally annotated proteins of *C. pneumoniae* and *C. trachomatis* from the COG search tool (Fig. 14.15). This shows proteins that are likely orthologs shared in common between these two organisms as well as proteins unique to one or the other species. There are hundreds of genes present uniquely in each bacterium, including a family of outer membrane proteins that could be important in tissue tropism (Kalman et al., 1999).

TaxPlot

The NCBI offers a powerful tool for genome comparison that is easy to use. From the Entrez Genome page, select *C. trachomatis* to obtain a page such as that shown in Figure 12.9. Select TaxPlot, and you will be able to compare two genomes (such as *C. trachomatis* and *C. pneumoniae* AR39) against a reference genome (the anthrax

TABLE 14-10 Prokaryotic Species for Which Genome of At Least Two Closely Related Strains Have Been Determined

Organism	Accession	Genome Size (bp)
Chlamydophila pneumoniae AR39	NC_002179	1,229,858
C. pneumoniae CWL029	NC_000922	1,230,230
C. pneumoniae J138	NC_002491	1,226,565
Escherichia coli K12	NC_000913	4,639,221
E. coli O157:H7	NC_002695	5,498,450
E. coli O157:H7 EDL933	NC_002655	5,528,445
Helicobacter pylori 26695	NC_000915	1,667,867
H. pylori J99	NC_000921	1,643,831
Mycobacterium tuberculosis CDC1551	NC_002755	4,403,836
M. tuberculosis H37Rv	NC_000962	4,411,529
Neisseria meningitidis MC58	NC_003112	2,272,351
N. meningitidis Z2491	NC_003116	2,184,406
Staphylococcus aureus aureus MW2	NC_003923	2,820,462
S. aureus aureus Mu50	NC_002758	2,878,040
S. aureus aureus N315	NC_002745	2,813,641
Streptococcus agalactiae 2603V/R	NC_004116	2,160,267
S. agalactiae NEM316	NC_004368	2,211,485
Streptococcus pneumoniae R6	NC_003098	2,038,615
S. pneumoniae TIGR4	NC_003028	2,160,837
Streptococcus pyogenes M1 GAS	NC_002737	1,852,441
S. pyogenes MGAS315	NC_004070	1,900,521
S. pyogenes MGAS8232	NC_003485	1,895,017

bacterium *B. anthracis* in the example of Fig. 14.16). In this plot, each point represents a protein in the reference genome. The x and y coordinates show the BLAST score for the closest match of each protein to the two *Chlamydia* proteomes being compared. Most proteins are found along a diagonal line, indicating that they have equal (or nearly equal) scores between the reference protein and either of the *Chlamydia* proteins. However, there are notable outliers, which could represent genes important in the distinctive behavior of these two organisms. These points are clickable (see circled data point in Fig. 14.16, arrow 3), and the selected data point is highlighted (Fig. 14.16, arrow 4). This protein is identified as an ABC transport protein, and there are further links to the pairwise BLAST comparisons (Fig. 14.16, arrow 5). The displays in TaxPlot can further be color coded according to the COG classification scheme.

TaxPlot is thus an easy way to identify proteins that are different in two microbial genomes of interest. The tool has been extended to eukaryotes as well (Chapter 16).

MUMmer

A major challenge in aligning whole microbial genomes is the excessive amount of time required to perform an alignment of millions of base pairs using dynamic

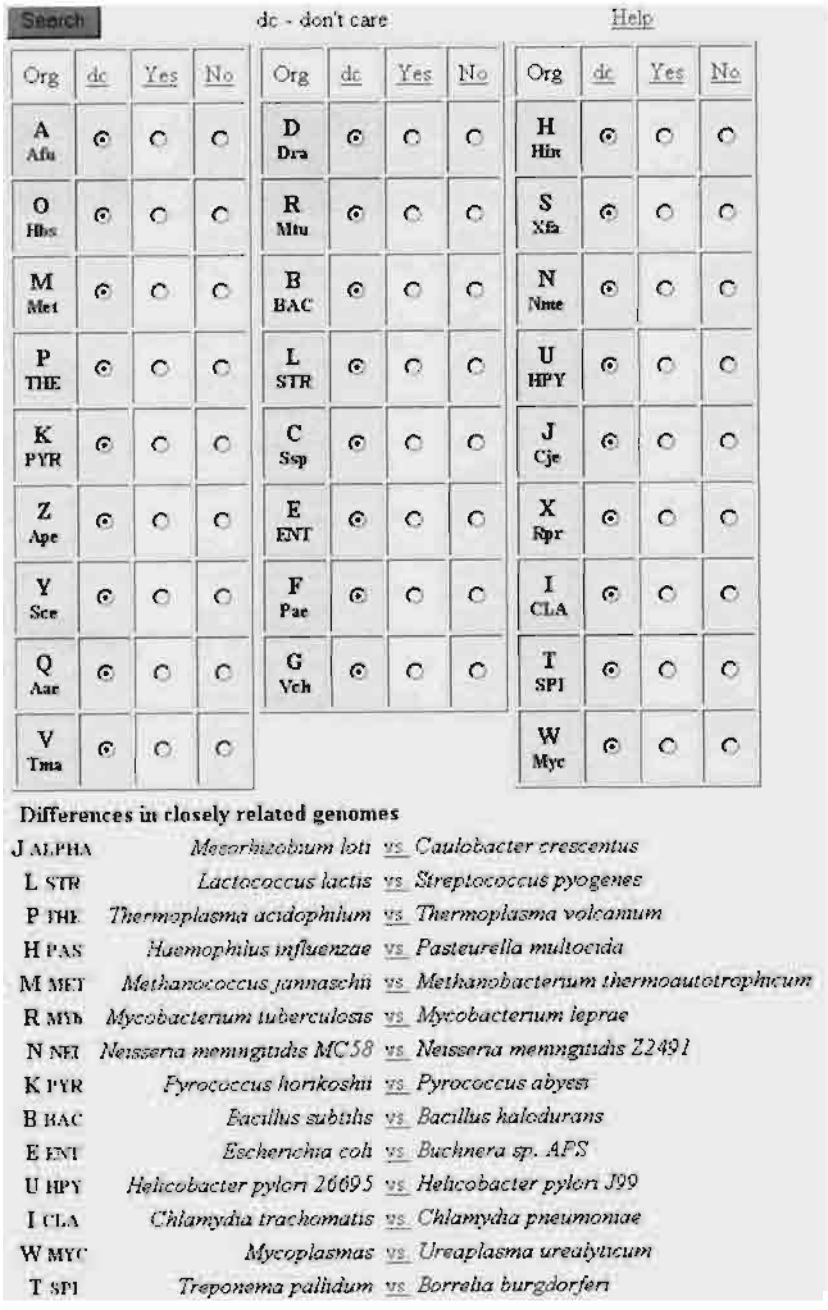


FIGURE 14.14. *The COG site allows the comparison of orthologous genes in specific genomes. Such analyses can help you to explore why one species of bacterium is pathogenic while a closely related species is not, or why an archeon is adapted to occupy a selected ecological niche.*

programming (Chapter 3). We introduced several fast algorithms such as BLAT in Chapter 5. Still, the tools to accomplish genome alignment are only beginning to be developed (Miller, 2001). MUMmer is a software package that offers an alternative approach to the rapid, accurate alignment of microbial genomes (Delcher et al., 1999b). In recent improvements to the algorithm, it is adapted to aligning eukaryotic sequences as well (Delcher et al., 2002).

MUMmer was written by Steven Salzberg and colleagues at TIGR. You can access it at ▸ http://www.tigr.org/software/mummer/, and there is an interactive web browser as well (▸ http://www.tigr.org/tigr-scripts/CMR2/webmum/mumplot).

MUMmer accepts two sequences as input. An example from the MUMmer web browser is shown in Figure 14.17. The algorithm finds all subsequences that are longer than a specified minimum length k and that are perfectly matched. By definition, these matches are maximal because extending them further in either direction causes a mismatch. The algorithm uses a suffix tree, which is a search structure that

Chalmydia trachomatis	-	x	-	x	
Chalmydia pneumoniae	-	-	x	x	
Function	2574	8	22	562	
J K L	328	2	3	185	**Information storage and processing**
J	99	-	2	112	Translation, ribosomal structure and biogenesis
K	100	1	-	23	Transcription
L	129	1	1	50	DNA replication, recombination and repair
D O M N P T	500	-	2	138	**Cellular processes**
D	24	-	-	7	Cell division and chromosome partitioning
O	81	-	-	25	Posttranslational modification, protein turnover, chaperones
M	113	-	1	34	Cell envelope biogenesis, outer membrane
N	85	-	-	36	Cell motility and secretion
P	129	-	1	22	Inorganic ion transport and metabolism
T	68	-	-	14	Signal transduction mechanisms
C G E F H I Q	770	5	14	189	**Metabolism**
C	184	1	-	35	Energy production and conversion
G	126	-	-	31	Carbohydrate transport and metabolism
E	179	3	5	40	Amino acid transport and metabolism
F	64	-	5	16	Nucleotide transport and metabolism
H	113	-	4	35	Coenzyme metabolism
I	48	1	-	26	Lipid metabolism
Q	56	-	-	6	Secondary metabolites biosynthesis, transport and catabolism
R S	1045	1	5	86	**Poorly characterized**
R	367	1	4	60	General function prediction only
S	678	-	1	26	Function unknown

FIGURE 14.15. *Using the COG phylogenetic pattern search tool, orthologous proteins in* C. trachomatis *and* C. pneumoniae *are compared. Both of these organisms are obligate intracellular pathogens.* Chlamydia pneumoniae *causes 10% of the pneumonia cases in the United States and 5% of the bronchitis cases, while* C. trachomatis *causes trachoma (an ocular infection that leads to blindness) and sexually transmitted diseases. A comparison of the proteins encoded by these two genomes could explain their distinct patterns of infection. This COG analysis shows many proteins that are predicted to be selectively present in one or the other of these bacteria.*

identifies all the maximal unique matches ("MUM"s) in the pairwise alignment. The MUMs are ordered, and the algorithm closes gaps by identifying large inserts, repeats, small mutated regions, and single-nucleotide polymorphisms (SNPs).

MUMmer output consists of a dot matrix plot (Fig. 14.18) showing the alignment of the two genomic sequences with some minimum alignment length (e.g., 15 or 100 bp). The kinds of results that can be obtained include:

1. SNPs
2. Regions where sequences diverge by more than a SNP
3. Large insertions (e.g., by transposition, sequence reversal, or lateral gene transfer)
4. Repeats (e.g., a duplication in one genome)
5. Tandem repeats (in different copy number)

In the example of Figure 14.18, two strains of *E. coli* are compared: a harmless *E. coli* K12 strain and the *E. coli* O157:H7 strain that appears in contaminated food,

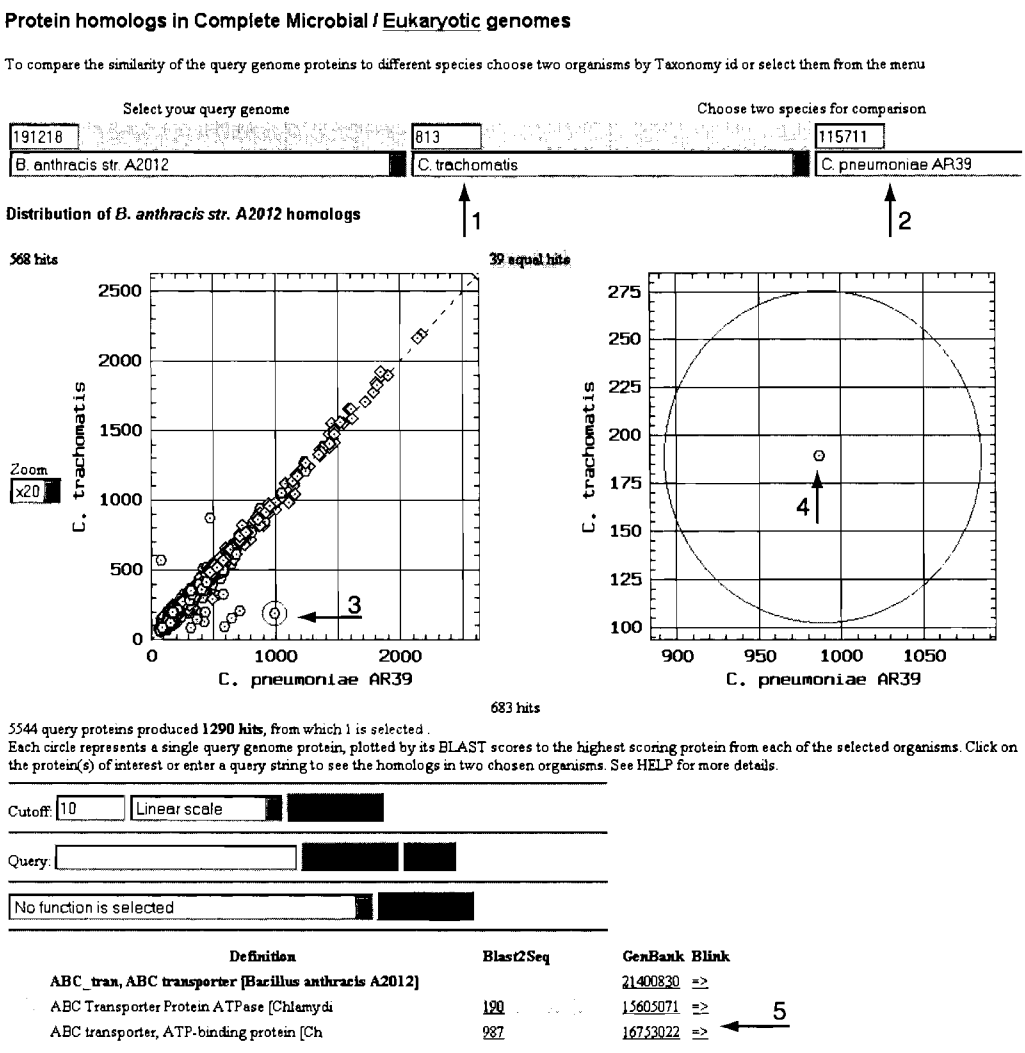


FIGURE 14.16. *The TaxPlot tool at NCBI (Entrez) allows the comparison of two bacteria (*C. trachomatis *and* C. pneumoniae *AR39) to a reference genome (*B. anthracis *strain A2012 in this case). The plot shows the distribution of BLASTP scores of each bacterium against the reference genome. Thirty-nine matches are identical, while about 600 hits are at least marginally closer to* C. trachomatis *or* C. pneumoniae.

causing disease such as hemorrhagic colitis. These strains diverged about 4.5 MYA (Reid et al., 2000). Both genomes were sequenced and compared (Blattner et al., 1997; Perna et al., 2001; Hayashi et al., 2001; reviewed in Eisen, 2001). *Escherichia coli* 0157:H7 is about 859,000 bp larger than *E. coli* K12. The two bacteria share a common backbone of about 4.1 Mb, while *E. coli* 0157:H7 has an additional 1.4-Mb sequence comprised largely of genes acquired by lateral gene transfer. The MUMmer output is useful to identify regions of the two genomes that are shared in common as well as regions in which the orientation is inverted.

While we have focused on TIGR and NCBI tools, the European Bioinformatics Institute also offers important microbial resources. ARTEMIS (Mural, 2000) is another tool that is useful for microbial genome alignment and analysis (▸http://www.sanger.ac.uk/Software/Artemis/).

PERSPECTIVE

The recent sequencing of several dozen bacterial and archaeal genomes has had a profound effect on virtually all aspects of microbiology. We can summarize the

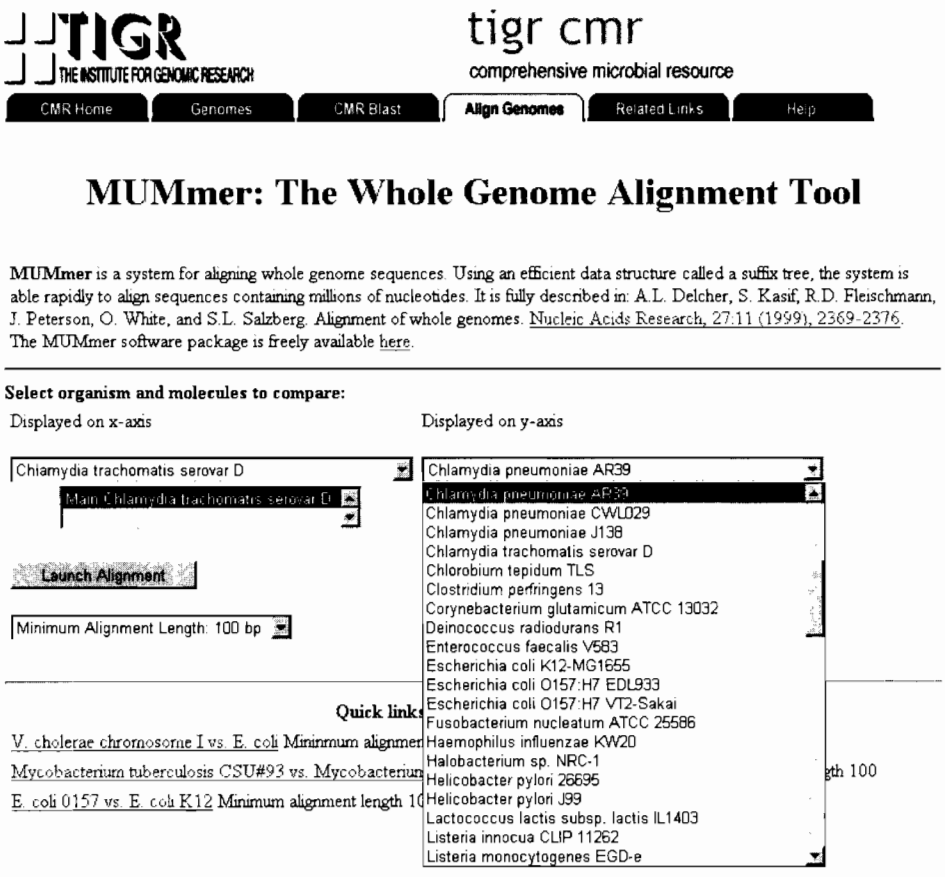


FIGURE 14.17. *A MUMmer web browser at TIGR allows you to select two microbial genomes of interest for comparison on a dot plot. The minimal alignment length can be adjusted.*

benefits of whole-genome sequencing of microbes as follows:

- Upon identifying the entire DNA sequence of a bacterial or archaeal genome, we obtain a comprehensive survey of all the genes and regulatory elements. This is similar to obtaining a parts list of a machine, although we do not also have the instruction manual.
- Through comparative genomics, we may learn the principles by which the "machine" is assembled and by which it functions.
- We can understand the diversity of microbial species through comparative genomics. Thus we can begin to uncover the principles of genome organization, and we can compare pathogenic versus nonpathogenic strains. We can also appreciate the dramatic differences in genome properties between two strains from the same species.
- We are gaining insights into the evolution of both genes and species. We can now appreciate lateral gene transfer as one of the driving forces of microbial evolution. We can study gene duplication and gene loss. Having the complete genome available is important both to learn what genes comprise an organism as well as to learn what genes are absent (Doolittle, 1998).
- Complete genome sequences offer a starting point for biological investigations.

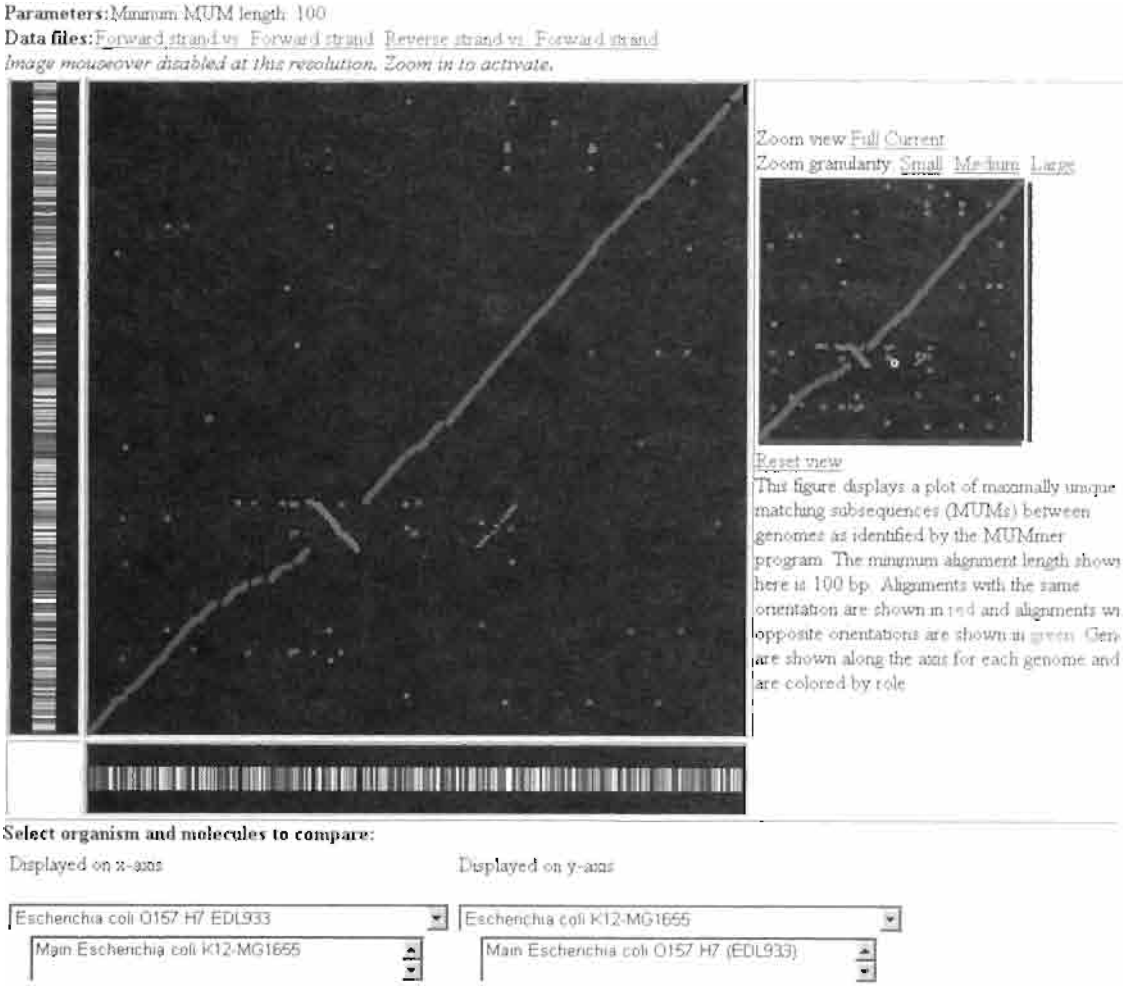


FIGURE 14.18. *The MUMmer output consists of a dot plot that displays maximally unique matching subsequences (MUMs) between two genomes. This tool rapidly describes the relationship between two genomes, including information on the relative orientation of the genomic DNA and the presence of insertions or deletions.*

PITFALLS

As complete bacterial and archaeal genomes are sequenced, two of the most important tasks are gene identification and genome annotation. Gene identification is difficult for several reasons: It is difficult to assess whether short ORFs correspond to transcripts that are actively transcribed, and (in contrast to eukaryotes) prokaryotes do not always use AUG as a start codon. Genome annotation is the critical process by which functions are assigned to predicted proteins. Computational annotation should always be viewed as a hypothesis that needs to be experimentally tested. There are several kinds of common errors (Brenner, 1999; Mural, 1999; Peri et al., 2001):

- Transitive catastrophes: inappropriately assigning a function to a gene based upon homology to another gene with a known function.
- Identification of small ORFs as authentic genes when they are not transcribed. Devos and Valencia (2001) estimate that about 5% of the genes annotated for general functions are incorrect, while about 33% of the gene annotations for specific functions are erroneous.

WEB RESOURCES

The TIGR Comprehensive Microbial Resource (▸http://www.tigr.org) provides an important starting point for any study of microbial genomes. That site includes a links page, ▸http://www.tigr.org/CMR2/Links.shtml, that provides a superb gateway to additional resources. A useful link on microbes is ▸http://www.microbes.info. This site includes a broad variety of resources including introductory articles on microbiology.

DISCUSSION QUESTIONS

[14-1] Anthrax strains vary in their pathogenicity. What bioinformatics approaches could you take to understand the basis of this difference?

[14-2] How can you assess whether bacterial genes have incorporated into the human genome through lateral gene transfer? What alternative explanations could there be for the presence of a human protein that is most closely related to a group of bacterial proteins, without having other eukaryotic orthologs?

[14-3] Consider the differences between *E. coli* K12 and *E. coli* O157:H7 and other closely related pairs of bacteria. They undergo lateral gene transfer to different degrees, they have distinct patterns of pathogenicity, and these two strains even differ in genome size by over a million base pairs. What is the definition of a species? Is *E. coli* a species?

PROBLEMS

[14-1] Analyze a bacterial genome. Select any bacterium for which the genome has been sequenced. If you cannot decide, take *E. coli*. Begin at Entrez Genomes. Find a gene that is known to have a homolog in eukaryotes. Use the TaxPlot tool of Entrez genomes. Now use the Clusters of Orthologous Genes (COG) site to find a gene that is known to have a homolog in eukaryotes. In addition to NCBI, there are two excellent resources for completed genomes:

- Explore this same bacterial genome at the TIGR website. Go to ▸http://www.tigr.org. Then (from the left sidebar) click TIGR databases. Click the link for "projects completed" and find your genome.
- Explore this same bacterial genome at The Wellcome Trust Sanger Institute website. Go to ▸http://www.sanger.ac.uk/Projects/.

[14-2] Compare two completed genomes. Begin at Entrez Genomes. Choose bacteria, then choose an organism such as *Rickettsia prowazekii*. Use TaxPlot to perform a three-way genome comparison. Try clicking on a point on the graph. Restrict your analysis to a functional group of genes ("transcription"). Repeat your search with the group "function unknown." Are the profiles different?

SELF-TEST QUIZ

[14-1] A typical bacterial genome is composed of approximately how many base pairs of DNA?
(a) 20,000 bp
(b) 200,000 bp
(c) 2,000,000 bp (2 Mb)
(d) 20,000,000 bp (20 Mb)

[14-2] In general, specialized bioinformatics tools are needed to study bacterial genomes because
(a) bacteria use their own genetic code, in contrast to eukarotes
(b) bacteria lack nuclei, in contrast to eukarotes
(c) bacteria have very few genes that are homologous to those in eukaryotes
(d) bacterial DNA has a gene content and organization that differs greatly from that of eukaryotes

[14-3] The *E. coli* genome encodes about 4,300 protein-coding genes. The total number of *E. coli* introns is approximately
(a) 10
(b) 430
(c) 4,300
(d) 43,000

[14-4] The smallest prokaryotic genomes tend to be those of
(a) extremophiles
(b) viruses
(c) intracellular bacteria
(d) Bacilli

[14-5] Which of the following options is best to determine how many *E. coli* proteins have orthologs in the proteome of a completely sequenced archaeon?

(a) PubMed
(b) ColiBase
(c) TaxPlot
(d) Mummer

[14-6] Which of the following constitutes strongest evidence that an *E. coli* gene became incorporated into the *E. coli* genome by lateral gene transfer?
(a) the GC content of the gene varies greatly relative to other *E. coli* genes
(b) the frequency of codon utilization of the gene varies greatly relative to other *E. coli* genes
(c) phylogenetic analysis shows that proteobacteria closely related to *E. coli* lack this gene
(d) any of the above

[14-7] The main idea of the Clusters of Orthologous Genes (COG) database is
(a) to classify proteins from completely sequenced prokaryotic genomes based on orthologous relationships
(b) to provide multiple sequence alignments of completed prokaryotic genomes
(c) to provide a functional classification system for proteins
(d) to predict the functions of individual eukaryotic proteins based on the conserved families in prokaryotes

Suggested Reading

There is a large literature on bacterial genomics. Casjens (1998) provides an excellent introduction. For an overview of the meaning of bacterial populations from a genomics perspective, see Joyce et al. (2002).

References

Achenbach-Richter, L., Gupta, R., Stetter, K. O., and Woese, C. R. Were the original eubacteria thermophiles? *Syst. Appl. Microbiol.* **9,** 34–39 (1987).

Akman, L., Rio, R. V., Beard, C. B., and Aksoy, S. Genome size determination and coding capacity of *Sodalis glossinidius,* an enteric symbiont of tsetse flies, as revealed by hybridization to *Escherichia coli* gene arrays. *J. Bacteriol.* **183,** 4517–4525 (2001).

Akman, L., Yamashita, A., Watanabe, H., Oshima, K., Shiba, T., Hattori, M., and Aksoy, S. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia. Nat Genet.* **32,** 402–407 (2002).

Alm, R. A., et al. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori. Nature* **397,** 176–180 (1999).

Andersson, S. G., and Kurland, C. G. Reductive evolution of resident genomes. *Trends Microbiol.* **6,** 263–268 (1998).

Andersson, S. G., et al. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396,** 133–140 (1998).

Baytaluk, M. V., Gelfand, M. S., and Mironov, A. A. Exact mapping of prokaryotic gene starts. *Brief. Bioinform.* **3,** 181–194 (2002).

Bender, L., Ott, M., Marre, R., and Hacker, J. Genome analysis of *Legionella* ssp. by orthogonal field alternation gel electrophoresis (OFAGE). *FEMS Microbiol. Lett.* **60,** 253–257 (1990).

Bentley, S. D., et al. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417,** 141–147 (2002).

Blattner, F. R., et al. The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1474 (1997).

Brenner, S. E. Errors in genome annotation. *Trends Genet.* **15,** 132–133 (1999).

Brochier, C., and Philippe, H. Phylogeny: A non-hyperthermophilic ancestor for bacteria. *Nature* **417,** 244 (2002).

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E., and Stanhope, M. J. Universal trees based on large combined protein sequence data sets. *Nat. Genet.* **28,** 281–285 (2001).

Bult, C. J., et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* **273,** 1058–1073 (1996).

Casjens, S. The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.* **32,** 339–377 (1998).

Casjens, S., et al. A bacterial genome in flux: The twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of the Lyme disease spirochete *Borrelia burgdorferi. Mol. Microbiol.* **35,** 490–516 (2000).

Chambaud, I., et al. The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis. Nucleic Acids Res.* **29,** 2145–2153 (2001).

Charon, N. W., and Goldstein, S. F. Genetics of motility and chemotaxis of a fascinating group of bacteria: The Spirochetes. *Annu. Rev. Genet.* **36,** 47–73 (2002).

Cole, S. T., et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393,** 537–544 (1998).

Cole, S. T., et al. Massive gene decay in the leprosy bacillus. *Nature* **409,** 1007–1011 (2001).

Dandekar, T., et al. Re-annotating the *Mycoplasma pneumoniae* genome sequence: Adding value, function and reading frames. *Nucleic Acids Res.* **28,** 3278–3288 (2000).

Deckert, G., et al. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392,** 353–358 (1998).

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27,** 4636–4641 (1999a).

Delcher, A. L., et al. Alignment of whole genomes. *Nucleic Acids Res.* **27,** 2369–2376 (1999b).

Delcher, A. L., Phillippy, A., Carlton, J., and Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30,** 2478–2483 (2002).

DeLong, E. F. Everything in moderation: Archaea as "non-extremophiles." *Curr. Opin. Genet. Dev.* **8,** 649–654 (1998).

DeLong, E. F., and Pace, N. R. Environmental diversity of bacteria and archaea. *Syst. Biol.* **50,** 470–478 (2001).

de Villiers, E. P., Brayton, K. A., Zweygarth, E., and Allsopp, B. A. Genome size and genetic map of *Cowdria ruminantium*. *Microbiology* **146,** 2627–2634 (2000).

Devos, D., and Valencia, A. Intrinsic errors in genome annotation. *Trends Genet.* **17,** 429–431 (2001).

Dobell, C. *Antony van Leeuwenhoek and his "little animals"*. Harcourt, Brace and Company, New York, 1932.

Doolittle, R. F. Microbial genomes opened up. *Nature* **392,** 339–342 (1998).

Eisen, J. A. Horizontal gene transfer among microbial genomes: New insights from complete genome analysis. *Curr. Opin. Genet. Dev.* **10,** 606–611 (2000).

Eisen, J. A. Gastrogenomics. *Nature* **409,** 463, 465–466 (2001).

Eisenstein, B. I., and Schaechter, M. Normal microbial flora. In M. Schaechter, N. C. Engleberg, B. I. Eisenstein, and G. Medoff (Eds.), *Mechanisms of Microbial Disease*. Lippincott Williams and Wilkins, Baltimore, MD, 1999, Chapter 20.

Ermolaeva, M. D., White, O., and Salzberg, S. L. Prediction of operons in microbial genomes. *Nucleic Acids Res.* **29,** 1216–1221 (2001).

Feng, D. F., Cho, G., and Doolittle, R. F. Determining divergence times with a protein clock: Update and reevaluation. *Proc. Natl. Acad. Sci. USA* **94,** 13028–13033 (1997).

Fleischmann, R. D., et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269,** 496–512 (1995).

Fox, G. E., et al. The phylogeny of prokaryotes. *Science* **209,** 457–463 (1980).

Fraser, C. M., et al. The minimal gene complement of *Mycoplasma genitalium*. *Science* **270,** 397–403 (1995).

Fraser, C. M., et al. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature* **390,** 580–586 (1997).

Fraser, C. M., et al. Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281,** 375–388 (1998).

Fraser, C. M., Eisen, J. A., and Salzberg, S. L. Microbial genome sequencing. *Nature* **406,** 799–803 (2000).

Glaser, P., et al. Comparative genomics of *Listeria* species. *Science* **294,** 849–852 (2001).

Glass, J. I., et al. The complete sequence of the mucosal pathogen *Ureaplasma urealyticum*. *Nature* **407,** 757–762 (2000).

Graur, D., and Li, W.-H. *Fundamentals of Molecular Evolution*. Sinauer Associates, Sunderland, MA, 2000, p. 481.

Gupta, R. S., and Griffiths, E. Critical issues in bacterial phylogeny. *Theor. Popul. Biol.* **61,** 423–434 (2002).

Hayashi, T., et al. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* **8,** 11–22 (2001).

Heidelberg, J. F., et al. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406,** 477–483 (2000).

Himmelreich, R., et al. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24,** 4420–4449 (1996).

Huber, H., et al. A new phylum of *Archaea* represented by a nanosized hyperthermophilic symbiont. *Nature* **417,** 63–67 (2002).

Hugenholtz, P. Exploring prokaryotic diversity in the genomic era. *Genome Biol.* **3**(2) (2002).

Hugenholtz, P., Goebel, B. M., and Pace, N. R. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180,** 4765–4774 (1998).

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

Itaya, M. An estimation of minimal genome size required for life. *FEBS Lett.* **362,** 257–260 (1995).

Joyce, E. A., Chan, K., Salama, N. R., and Falkow, S. Redefining bacterial populations: A post-genomic reformation. *Nat. Rev. Genet.* **3,** 462–473 (2002).

Kalman, S., et al. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat. Genet.* **21,** 385–389 (1999).

Kawarabayasi, Y., et al. Complete genome sequence of an aerobic hyperthermophilic crenarchaeon, *Aeropyrum pernix* K1. *DNA Res.* **6,** 83–101, 145–152 (1999).

Klenk, H. P., et al. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature* **390,** 364–370 (1997).

Koonin, E. V. Genome sequences: Genome sequence of a model prokaryote. *Curr. Biol.* **7,** R656–R659 (1997).

Koonin, E. V., Makarova, K. S., and Aravind, L. Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu. Rev. Microbiol.* **55,** 709–742 (2001).

Kunst, F., et al. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis. Nature* **390,** 249–256 (1997).

Ma, J., Campbell, A., and Karlin, S. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J. Bacteriol.* **184,** 5733–5745 (2002).

Malloff, C. A., Fernandez, R. C., Dullaghan, E. M., Stokes, R. W., and Lam, W. L. Two-dimensional display and whole genome comparison of bacterial pathogen genomes of high G+C DNA content. *Gene* **293,** 205–211 (2002).

May, B. J., et al. Complete genomic sequence of *Pasteurella multocida,* Pm70. *Proc. Natl. Acad. Sci. USA* **98,** 3460–3465 (2001).

Miller, W. Comparison of genomic DNA sequences: Solved and unsolved problems. *Bioinformatics* **17,** 391–397 (2001).

Mural, R. J. Current status of computational gene finding: a perspective. *Methods Enzymol.* **303,** 77–83 (1999).

Mural, R. J. ARTEMIS: A tool for displaying and annotating DNA sequence. *Brief. Bioinform.* **1,** 199–200 (2000).

Mushegian, A. R., and Koonin, E. V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA* **93,** 10268–10273 (1996).

Nelson, K. E., et al. Evidence for lateral gene transfer between *Archaea* and bacteria from genome sequence of *Thermotoga maritima. Nature* **399,** 323–329 (1999).

Nolling, J., et al. Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum. J. Bacteriol.* **183,** 4823–4838 (2001).

Ogata, H., et al. Mechanisms of evolution in *Rickettsia conorii* and *R. prowazekii. Science* **293,** 2093–2098 (2001).

Owen, Richard. Lectures on the Comparative Anatomy and Physiology of the Invertebrate Animals. Longman, Brown, Green, and Longmans, London, 1843.

Parkhill, J., et al. Complete genome sequence of a multiple drug resistant *Salmonella enterica* serovar *typhi* CT18. *Nature* **413,** 848–852 (2001a).

Parkhill, J., et al. Genome sequence of *Yersinia pestis,* the causative agent of plague. *Nature* **413,** 523–527 (2001b).

Peri, S., Ibarrola, N., Blagoev, B., Mann, M., and Pandey, A. Common pitfalls in bioinformatics-based analyses: Look before you leap. *Trends Genet.* **17,** 541–545 (2001).

Perna, N. T., et al. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409,** 529–533 (2001).

Perrière, G., Duret, L., and Gouy, M. HOBACGEN: Database system for comparative genomics in bacteria. *Genome Res.* **10,** 379–385 (2000).

Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K., and White, O. The Comprehensive Microbial Resource. *Nucleic Acids Res.* **29,** 123–125 (2001).

Ponting, C. P. Plagiarized bacterial genes in the human book of life. *Trends Genet.* **17,** 235–237 (2001).

Read, T. D., et al. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Res.* **28,** 1397–1406 (2000).

Read, T. D., et al. Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis. Science* **296,** 2028–2033 (2002).

Reid, S. D., Herbelin, C. J., Bumbaugh, A. C., Selander, R. K., and Whittam, T. S. Parallel evolution of virulence in pathogenic *Escherichia coli. Nature* **406,** 64–67 (2000).

Reysenbach, A. L., and Shock, E. Merging genomes with geochemistry in hydrothermal ecosystems. *Science* **296,** 1077–1082 (2002).

Salzberg, S. L., White, O., Peterson, J., and Eisen, J. A. Microbial genes in the human genome: Lateral transfer or gene loss? *Science* **292,** 1903–1906 (2001).

Schaechter, M. Introduction to the pathogenic bacteria. In M., Schaechter, N. C., Engleberg, B. I., Eisenstein, and G., Medoff (Eds.), *Mechanisms of Microbial Disease.* Lippincott Williams and Wilkins, Baltimore, MD, 1999, Chapter 10.

Serres, M. H., et al. A functional update of the *Escherichia coli* K-12 genome. *Genome Biol.* **2(9)**:research 0035.1–0035.7 (2001).

Sherratt, D. Divide and rule: The bacterial chromosome. *Trends Genet.* **17,** 312–313 (2001).

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y., and Ishikawa, H. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407,** 81–86 (2000).

Shirai, M., et al. Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. *Nucleic Acids Res.* **28,** 2311–2314 (2000).

Skovgaard, M., Jensen, L. J., Brunak, S., Ussery, D., and Krogh, A. On the total number of genes and their length distribution in complete microbial genomes. *Trends Genet.* **17,** 425–428 (2001).

Stanhope, M. J., et al. Phylogenetic analyses do not support horizontal gene transfers from bacteria to vertebrates. *Nature* **411,** 940–944 (2001).

Stephens, R. S., et al. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis. Science* **282,** 754–759 (1998).

Stover, C. K., et al. Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406,** 959–964 (2000).

Sun, L. V., et al. Determination of *Wolbachia* genome size by pulsed-field gel electrophoresis. *J. Bacteriol.* **183,** 2219–2225 (2001).

Tamas, I., et al. 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296,** 2376–2379 (2002).

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. A genomic perspective on protein families. *Science* **278,** 631–637 (1997).

Tatusov, R. L., et al. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29,** 22–28 (2001).

Tettelin, H., et al. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287,** 1809–1815 (2000).

Tomb, J. F., et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori. Nature* **388,** 539–547 (1997).

White, O., et al. Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286,** 1571–1577 (1999).

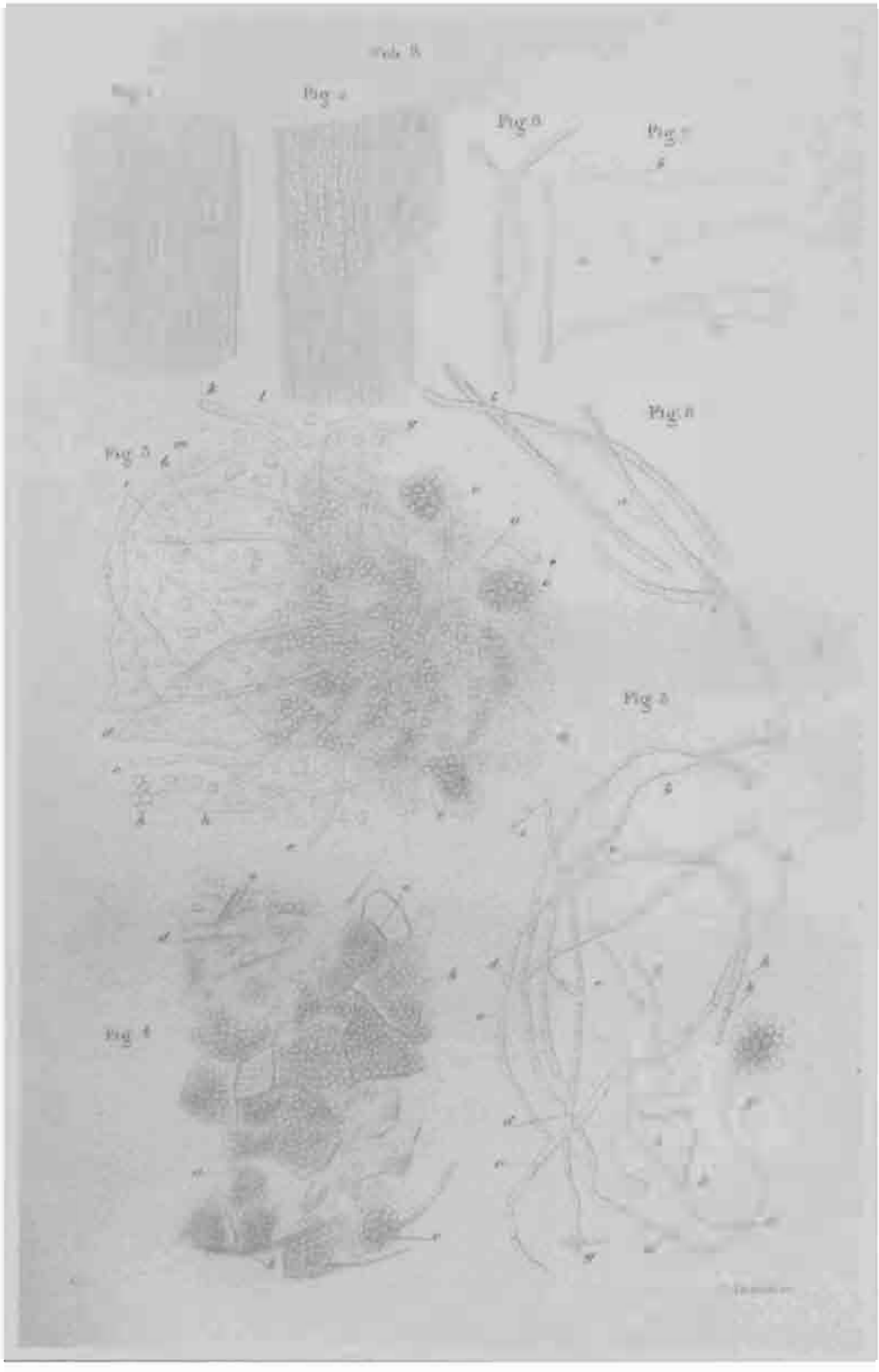
Woese, C. R., and Fox, G. E. The concept of cellular evolution. *J. Mol. Evol.* **10,** 1–6 (1977).

Woese, C. R. On the evolution of cells. *Proc. Natl. Acad. Sci. USA* **99,** 8742–8747 (2002).

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L., and Koonin, E. V. Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol. Biol.* **1,** 8 (2001).

Wood, V., et al. The genome sequence of *Schizosaccharomyces pombe. Nature* **415,** 871–880 (2002).

This Page Intentionally Left Blank



Some 200 fungal species are known to be pathogenic for humans, distressing millions of people. From the times of Greek and Roman antiquity to the middle of the 19th century only two fungal diseases were known: ringworm (tinea) and thrush (oral candidiasis) (Ainsworth, 1993). Ringworm is caused by fungi of the genera Microsporum, Trichophyton, *and* Epidermophyton. *Candidiasis (including thrush) is caused by* Candida albicans *and other* Candida species. *This image from Kuchenmeister (1857, plate IV) shows the thrush fungus, at that time called* Oidium albicans *(Fig. 3 to 8).*

15

Eukaryotic Genomes: Fungi

INTRODUCTION

According to Whittaker's classification system, there are five kingdoms of life: monera (prokaryotes), protoctists, animals, fungi, and plants. We have examined the prokaryotes in Chapter 14. In this chapter we begin our exploration of eukaryotes by studying one of the kingdoms, Fungi. This diverse and interesting group of organisms last shared a common ancestor with plants and animals 1.5 billion years ago (BYA) (Wang et al., 1999, discussed in Chapter 16). We may think of fungi as organisms such as mushrooms that might be studied by botanists. Surprisingly, fungi are far more closely related to animals than to plants. In Chapter 16 we will extend our study to the entire kingdom of eukaryotes, including animals, plants, and a variety of protozoa. We will then discuss humans (Chapter 17).

We begin our study of eukaryotes with the fungi. The first eukaryotic genome to be fully sequenced was the 13-million-base-pair (Mb) genome of a fungus, the budding yeast *Saccharomyces cerevisiae*. Its genome is very small compared with that of humans [3 billion base pairs, or gigabase pairs (Gb)], and its size is only severalfold larger than a typical bacterial genome. This yeast has served as a model eukaryotic organism for genetics studies because it grows rapidly, it can be genetically modified easily, and many of its cellular functions are conserved with metazoans and other eukaryotes. More recently, it has become a model organism for genomics and bioinformatics studies. Every one of its approximately 6000 genes has been characterized, deleted, overexpressed, and characterized functionally using a variety of assays (see below).

Mycology (from the Greek word *mukes*, "fungus") is the study of fungi. Mycosis is a disease or ailment caused by fungi. The suffix *-mycota* refers to fungi: The kingdom Fungi is also called the kingdom Eumycota ("true fungi").

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

This chapter begins with an overview of the fungi. We will then describe bioinformatic approaches to the *S. cerevisiae* genome. Finally, we will describe the sequencing of other fungal genomes and the early lessons of comparative genomics in fungi.

Description and Classification of Fungi

Fungi are eukaryotic organisms that can be filamentous (as in the case of molds) or unicellular (as in the case of yeasts such as *S. cerevisiae*). The main criteria for classifying fungi are based on morphology (e.g., ultrastructure), biochemistry (e.g., growth properties or cell wall composition), and molecular sequence data (DNA, RNA, and protein sequences). Most fungi are aerobic, and all are heterotrophs that absorb their food. Fungi are typically very hardy, forming spores composed of chitin that are immobile throughout their lifespan. They have a major role in the ecosystem in degrading organic waste material. Fungi are important causative agents of disease to humans, other animals, and plants. Fungi also have key roles in fermentation; the fungal mold *Rhizopus nigricans* is used in the manufacture of steroids such as cortisone, and *Penicillium chrysogenum* produces the antibiotic penicillin.

Morphologically, fungi are characterized by hyphae (filaments) that grow and may branch. The Museum of Paleontology at the University of California, Berkeley, offers an introduction to fungi, including photographs of many species (▸http://www.ucmp.berkeley.edu/fungi/fungi.html). The American Museum of Natural History (New York) also provides an overview of fungi (▸http://ology.amnh.org/biodiversity/treeoflife/pages/fungi.html).

The relationships of many species throughout the tree of life have been described in phylogenetic analyses based on small-subunit ribosomal RNA (Fig. 12.1). In a complementary approach, W. F. Doolittle and colleagues defined a phylogeny of the eukaryotes based upon the concatenated amino acid sequences from four proteins: elongation factor-1α, actin, α-tubulin, and β-tubulin (Baldauf et al., 2000). A portion of the tree shows that fungi form a monophyletic clade that is a sister group to animals (metazoa) (Fig. 15.1). This close relationship between fungi and animals has been considered somewhat surprising, given the apparently simple, unicellular nature of many fungi. However, fungi and animals share many similarities. Chitin is the main component of the fungal cell wall, and it is also a constituent of the arthropod exoskeleton. (Plant cell walls use cellulose.) Many of the fundamental processes of yeast, such as cell cycle control, DNA repair, and intracellular vesicle trafficking, are closely conserved with mammalian cells.

We will explore this comprehensive tree in detail in Chapter 16 (Fig. 16.1).

See Box 15.1 for a discussion of fungal taxonomy.

FIGURE 15.1. *Phylogenetic analysis of the fungi reveals that they form a sister group with the metazoa (animals). This tree is a detailed view of a broad analysis of the eukaryotes (see Fig. 16.1) by Baldauf et al. (2000). The tree was generated using a multiple sequence alignment of four concatenated protein sequences: elongation factor 1α (EF-1α) (abbreviated E in tree), actin (C), α-tubulin (A), and β-tubulin (B). Microsporidia were formerly classified as deep-branching eukaryotes but are now grouped with fungi. The fungal phylum Chytridiomycota is not shown in this tree.*

BOX 15-1
Fungal Taxonomy

Approximately 70,000 fungal species were described in 1995, although the total number of species is estimated to be at least 1.5 million. These fungi are classified in four phyla: Ascomycota, Basidiomycota, Chytridiomycota, and Zygomycota (Guarro et al., 1999). Ascomycota includes yeasts, blue-green molds, truffles, and lichens; about 30,000 species are known, including the genera *Aspergillus*, *Candida*, *Cryptosporium*, *Histoplasma*, *Neurospora*, and *Saccharomyces*. Basidiomycota includes rusts, smuts, and mushrooms; they are distinguished by club-shaped reproductive structures called basidia. The phylum Chytridiomycota, sometimes classified in the kingdom Protoctista (Margulis and Schwartz, 1998), includes the genera *Allomyces* and *Polyphagus*. Finally, fungi of the phylum Zygomycota lack septa (cross walls), typically feed on decaying vegetation and include the genera *Glomus*, *Mucor*, and *Rhizopus*.

The phylum Ascomycota is of particular interest because it includes the yeasts. The phylum is further divided into four classes: Hemiascomycetae (e.g., *S. cerevisiae*), Euascomycetae (e.g., *Neurospora crassa*), Loculoascomycetae (e.g., *Elsinoe proteae*), and Laboulbeniomycetae (parasites of insects).

In addition to the four phyla of the kingdom Fungi, some fungi are found in the kingdom Chromista (phyla Hyphochytriomycota, Labyrinthulomycota, and Oomycota) and in the kingdom Protozoa (phyla Acrasiomycota, Dictyosteliomycota, Myxomycota, and Plasmodiophoromycota). The presence of fungi in various kingdoms highlights the idea that taxonomy has historically relied on morphological criteria that are sometimes not consistent with molecular criteria.

Before the Apollo moon missions, some of the astronauts were infected with fungi, including *Aspergillus fumigatus*, *Candida albicans*, and *Trichophyton rubrum* (▸http://lsda.jsc.nasa.gov/scripts/cf/exper.cfm?exp_index = 372). In the closed environment of the command module, some of these pathogenic species lacked normal microbial competition and flourished during the space missions. Identifying and treating fungal infections is of concern for long space missions.

Fungi are grown on food products such as Camembert and Brie cheeses to provide flavor. Fungi are used to produce soy sauce and many other foods.

INTRODUCTION TO BUDDING YEAST *SACCHAROMYCES CEREVISIAE*

The budding yeast *S. cerevisiae* was the first species domesticated by humans at least 10,000 years ago. It is commonly called brewer's yeast or baker's yeast, and it ferments glucose to ethanol and carbon dioxide. For more than 100 years, researchers have exploited this organism for biochemical, genetic, molecular, and cell biological studies. Because many of its characteristics are conserved also in human cells, yeast has emerged as a powerful instrument for basic research.

From the time of Anton van Leeuwenhoek (1632–1723), yeast were thought to be chemical substances that are not living. Theodor Schwann (1810–1882) and Baron Charles Cagniard-Latour (1777–1859) independently discovered in 1836–1837 that yeast are composed of living cells. Schwann studied fermenting yeast and called them *Zuckerpilz* (sugar fungus), from which the term *Saccharomyces* is derived (Bulloch, 1938).

Saccharomyces cerevisiae is often called a "budding yeast" to distinguish it from a "fission yeast," *Schizosaccharomyces pombe*, the second fungal genome to be sequenced (see below). *Saccharomyces cerevisiae* is a single-celled organism that "buds" off in the process of replication.

The sequence of *S. cerevisiae* was first released April 24, 1996.

Sequencing Yeast Genome

Currently, relatively small genomes are sequenced using the whole-genome shotgun method (Chapter 12). In contrast, the yeast genome was sequenced in the early to mid-1990s by chromosome. This was accomplished by a consortium of over 600 researchers from around the world (Mewes et al., 1997). The work proceeded in several phases. First, a crude physical map of its 16 chromosomes was constructed using rare-cutter restriction enzymes. Second, a library of ~10-kb genomic DNA inserts was constructed in phage lambda, and the inserts were fingerprinted using restriction enzymes. Computer analysis identified clones with overlapping inserts, which were then assembled into 16 large contigs. A set of clones covering the genome with minimal overlap was selected and parsed out to individual laboratories for sequencing by various methods. Individuals were responsible for accuracy. Finally, the individual sequences were collected, compiled centrally, assembled into the 16 complete chromosomes, and annotated using a standardized nomenclature. [The final

TABLE 15-1 Features of *S. cerevisiae* Genome as Initially Annotated

Feature	Amount
Sequenced length	12,068 kb
Length of repeats	1321 kb
Total length	13,389 kb
ORFs	6,275
Questionable proteins	390
Hypothetical proteins	5885
Introns in ORFs	220
Introns in UTRs	15
Intact Ty elements	52
tRNA genes	275
snoRNA genes	40

Source: (Adapted from Goffeau et al., 1996). An updated version of this table is available at the MIPS Comprehensive Yeast Genome Database (▸http://mips.gsf.de/proj/yeast/CYGD/db/index.html).

Abbreviations: ORF, open reading frame; snoRNA, small nucleolar RNA; tRNA, transfer RNA; Ty, retrotransposons; UTR, untranslated region.

error rate was less than 3 per 10,000 bases, or 0.03% (Mewes et al., 1997).] Today, this approach would be considered arduous, inefficient, and expensive. However, the collaboration worked extremely well.

Features of Budding Yeast Genome

The *S. cerevisiae* genome consists of about 13 Mb of DNA in 16 chromosomes. With the complete sequencing of the genome, the physical map (determined directly from DNA sequencing) was unified with the genetic map (determined by tetrad analysis to derive genetic distances between genes) (Cherry et al., 1997). The final sequence was assembled from 300,000 independent sequence reads (Mewes et al., 1997). Some of the features of the *S. cerevisiae* sequence are listed in Table 15.1, based on the initial annotation of the genome (Goffeau et al., 1996).

A notable feature of the yeast genome is its high gene density (about one gene every 2 kb). While bacteria have a density of about one gene per kilobase, most higher eukaryotes have a much sparser density of genes. Also, only 4% of the genes are interrupted by introns. In contrast, in the fission yeast *S. pombe,* 40% of the genes have introns (see below). The lack of introns makes *S. cerevisiae* an attractive model organism for the identification of genes from genomic DNA. The most common protein families and protein domains in *S. cerevisiae* are listed in Tables 15.2 and 15.3. The EBI offers a variety of proteomics analyses of this and dozens of other organisms (Fig. 15.2), such as an analysis of protein lengths (Fig. 15.3).

At the time the genomic sequence was initially annotated, there were 6275 predicted open reading frames (ORFs). An ORF was defined as $\geq$100 codons (300 nucleotides) in length, thus specifying a protein of at least $\approx$11,500 daltons. Of these, 390 were listed as questionable (Table 15.1) because they were short and unlikely to encode proteins (Dujon et al., 1994). Questionable ORFs display an unlikely preference for codon usage based on a "codon adaptation index" of <0.11.

By definition, all ORFs begin with a start codon (typically AUG encoding methionine) and end with a stop codon (usually UAG, UAA, or UGA).

TABLE 15-2 Fifteen Most Common Protein Families for *S. cerevisiae*

InterPro ID	Number of Proteins Matched	Name
IPR000719	115	Protein kinase
IPR005828	53	General substrate transporter
IPR001042	52	TYA transposon protein
IPR001969	40	Eukaryotic/viral aspartic protease, active site
IPR000379	37	Esterase/lipase/thioesterase, active site
IPR001806	34	Ras GTPase superfamily
IPR001993	34	Mitochondrial substrate carrier
IPR005834	28	Haloacid dehalogenase-like hydrolase
IPR000992	27	Stress-induced protein SRP1/TIP1 family
IPR002293	25	Amino acid/polyamine transporter, family I
IPR001142	23	Yeast membrane protein DUP
IPR002085	21	Zinc-containing alcohol dehydrogenase superfamily
IPR001394	18	Ubiquitin thiolesterase, family 2
IPR001757	17	ATPase, E1-E2 type
IPR001163	16	Small nuclear ribonucleoprotein (Sm protein)

Source: From ▸http://www.ebi.ac.uk/proteome/, November 2002.

TABLE 15-3 Fifteen Most Common Domains of Completed Genome of Yeast *S. cerevisiae*

InterPro ID	Matches per Genome	Number of Proteins	Name
IPR001650	144	71	Helicase C-terminal domain
IPR003593	70	58	AAA ATPase
IPR001138	238	57	Fungal transcriptional regulatory protein, N terminus
IPR000504	301	55	RNA-binding region RNP-1 (RNA recognition motif)
IPR000822	378	53	Zinc finger, C2H2 type
IPR000822	43	42	Small GTP-binding protein domain
IPR001841	105	40	Zincfinger, RING
IPR001584	37	37	Integrase, catalytic domain
IPR001849	77	29	Pleckstrin homology (PH) domain
IPR001452	144	25	SH3 domain
IPR004841	22	22	Domain found in permeases
IPR001623	79	21	Heat shock protein DnaJ, N terminal
IPR004843	41	21	Metallo-phosphoesterase
IPR000051	20	20	SAM (and some other nucleotide) binding motif
IPR000629	20	20	ATP-dependent helicase, DEAD-box

Source: From the European Bioinformatics Institute (EBI) proteome analysis site (▶ http://www.ebi.ac.uk/proteome/, November 2002).

How many protein-coding genes are present in *S. cerevisiae*? Is it possible that short ORFs encode authentic proteins? These questions are fundamental to our understanding of any eukaryotic genome. In annotating the yeast genome, there are false positives (identified ORFs that do not encode an authentic gene) and false negatives (true genes with short ORFs that are not annotated). There are 40,000 ORFs longer than 20 codons (Mackiewicz et al., 2002). Below the arbitrary cutoff of 100 codons, there are many ORFs that meet the criteria of having a codon adaptation index of >0.11 and which do not overlap a longer ORF (Harrison et al., 2002). The main criteria for deciding whether they are protein-coding genes are (1) evidence of conservation in other organisms and/or (2) experimental evidence of gene expression (Chapter 12). For *S. cerevisiae*, Winzeler and colleagues used a combination of gene expression profiling with oligonucleotide arrays and mass spectrometry to verify the transcription of 138 and the translation of 50 previously nonannotated genes (Oshiro et al., 2002). Michael Snyder and colleagues combined expression profiling, transposon-mediated gene trapping (see below), and homology searching to identify 137 genes (Kumar et al., 2002a).

In addition to protein-coding genes, there are many transcribed genes that encode functional RNA molecules but are not subsequently translated into protein. In addition to the 275 tRNA genes shown in Table 15.1, there are 140 tandemly repeated copies of rRNA genes as well as small nucleolar (snoRNA) (Lowe and Eddy, 1999) and other RNA species.

The MIPS Comprehensive Yeast Genome Database is a central repository for information on the *S. cerevisiae* genome (Mewes et al., 2002). This database describes ORFs in several categories based on FASTA scores (Table 15.4). Proteins with "weak

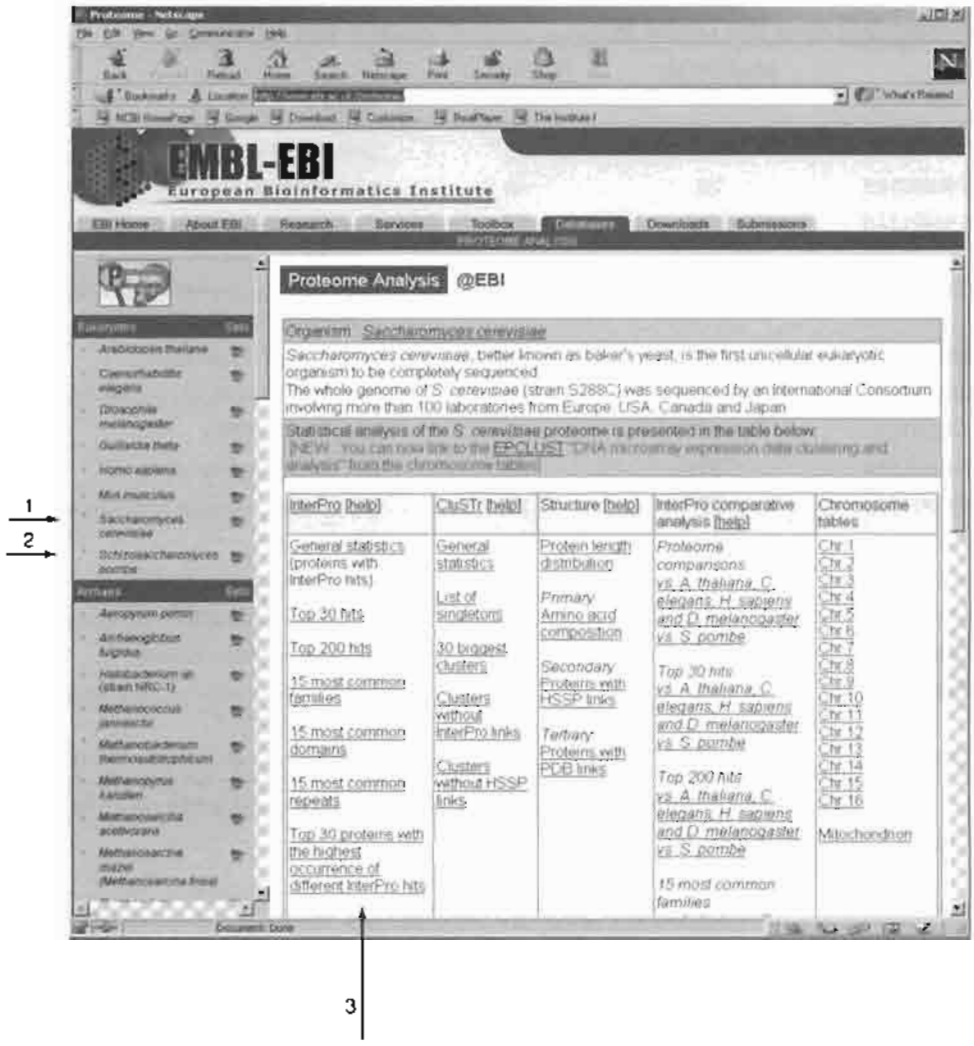


FIGURE 15.2. The European Bioinformatics Institute offers proteome analysis tools for S. cerevisiae *and dozens of other organisms (▸http://www.ebi.ac.uk/proteome/). Select the organism of choice from the left sidebar (arrows 1 and 2 indicate two fungal proteomes). Dozens of links to the selected proteome are provided, including InterPro resources (arrow 3).*

similarity" have a FASTA score between 100 and 200. Proteins with "strong similarity" have a FASTA score between 200 and one-third of the "self-score" (the score of the protein aligned against itself). A questionable ORF has a low codon adaptation index, partial overlap to a longer or known ORF, and/or no similarity to other ORFs.

The *S. cerevisiae* genome encodes 52 intact retrotransposons (called Ty1, Ty2, Ty3, Ty4, and Ty5). These are endogenous retrovirus-like elements that mediate transposition (i.e., insertion into a new genomic location) (Roth, 2000). They are flanked by long-terminal repeats (LTRs) that function in integration of the retrotransposon into a new genomic site. Retrotransposons have shaped the genomic landscape of all eukaryotic genomes.

Over half the human genome is composed of transposable elements; we will explore them in more detail in Chapters 16 and 17.

Exploring Typical Yeast Chromosome

You can access the DNA sequence of any *S. cerevisiae* chromosome through several websites. We begin with the Entrez Genomes division of NCBI; click "prominent organisms" on the left sidebar and then select any of the 16 chromosomes (Fig. 15.4). Consider the specific features of chromosome XII, a typical chromosome consisting of just over 1 Mb (Johnston et al., 1997) (Fig. 15.5). The NCBI page for this chromosome offers features such as TaxPlot and COG that are similar to the

Chromosome XII (accession number NC_001144) has 1,078,173 bp. To view a map of chromosome XII at the MIPS site, go to ▸http://mips.gsf.de/proj/yeast/CYGD/db/index.html. For SGD, visit ▸http://genome-www.stanford.edu/Saccharomyces/MAP/GENOMICVIEW/GenomicView.html. We will explore these sites below in more detail.

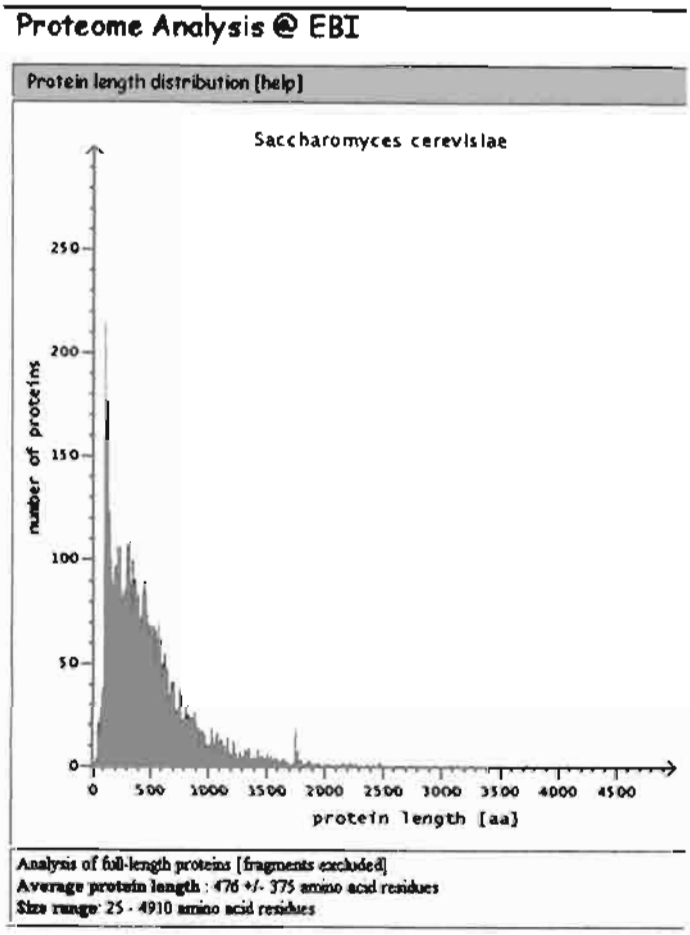


FIGURE 15.3. *The EBI proteome analysis includes a plot of the number of* S. cerevisiae *proteins as a function of the length of the protein. In the case of small predicted proteins (e.g., <100 codons) it is important to confirm that the gene is transcribed and translated in vivo and does not represent a fortuitous open reading frame that is not biologically meaningful.*

features we have seen for prokaryotes (Chapter 14). You can also view chromosome XII at the major yeast-specific databases such as the MIPS Comprehensive Yeast Genome Database (Fig. 15.6) and the *Saccharomyces* Genome Database (SGD) (Dwight et al., 2002) (Fig. 15.7):

- The overall G+C content of chromosome XII is 38%. The G+C content tends to be highest in localized regions corresponding to a high density of protein-coding genes. There are three regions of particularly low G+C content (below 37%); one of these corresponds to the centromere. This feature is typical of all eukaryotic centromeres.
- Overall, there is very little repetitive DNA throughout the *S. cerevisiae* genome. There are rDNA repeats on chromosome XII (encoding rRNAs). This region of the chromosome has the highest G+C content as well (approximately 42%). In addition, *S. cerevisiae* chromosomes have telomeric and subtelomeric repetitive DNA elements. This feature is typical of essentially all eukaryotic chromosomes.
- There are few spliceosomal introns (~235 total). These are probably due to homologous recombination of cDNAs produced by reverse transcription of spliced mRNAs. On chromosome XII, 17 ORFs (3.2% of the total) contain introns; half of these genes encode ribosomal proteins.

The centromere is the site at which chromosomes attach to the mitotic or meiotic spindle. In yeast, the centromere divides each chromosome into the left and right arm; in humans, it divides each chromosome into a short (or p) arm and a long (or q) arm.

The telomere is the terminal region of each chromosome arm. These arms are important in the maintenance of chromosome structure. They have been implicated in processes ranging from aging to mental retardation (Chapter 18).

TABLE 15-4 Categories of Yeast ORFs from Munich Information Center for Protein Sequences (MIPS) Comprehensive Yeast Genome Database

This database provides full lists of each of these categories of ORFs	
Category	Number of Proteins
Known protein	3400
Strong similarity to known protein	230
Similarity or weak similarity to known protein	825
Similarity to unknown protein	1007
No similarity	516
Questionable ORF	472
Total	6450

Source: From ▸ http://mips.gsf.de/proj/yeast/CYGD/db/index.html, November, 2002.

- There are six transposable elements (Ty elements) on chromsome XII. Additionally there are hundreds of fragments of transposable elements.
- The density of ORFs is extremely high. Seventy-two percent of chromosome XII contains protein-coding genes, a fraction that is typical of the other yeast

FIGURE 15.4. The NCBI Entrez Genomes site includes this page on S. cerevisiae. Each of the 16 chromosomes can be explored separately. The left sidebar includes links to major web resources for S. cerevisiae and other fungi.

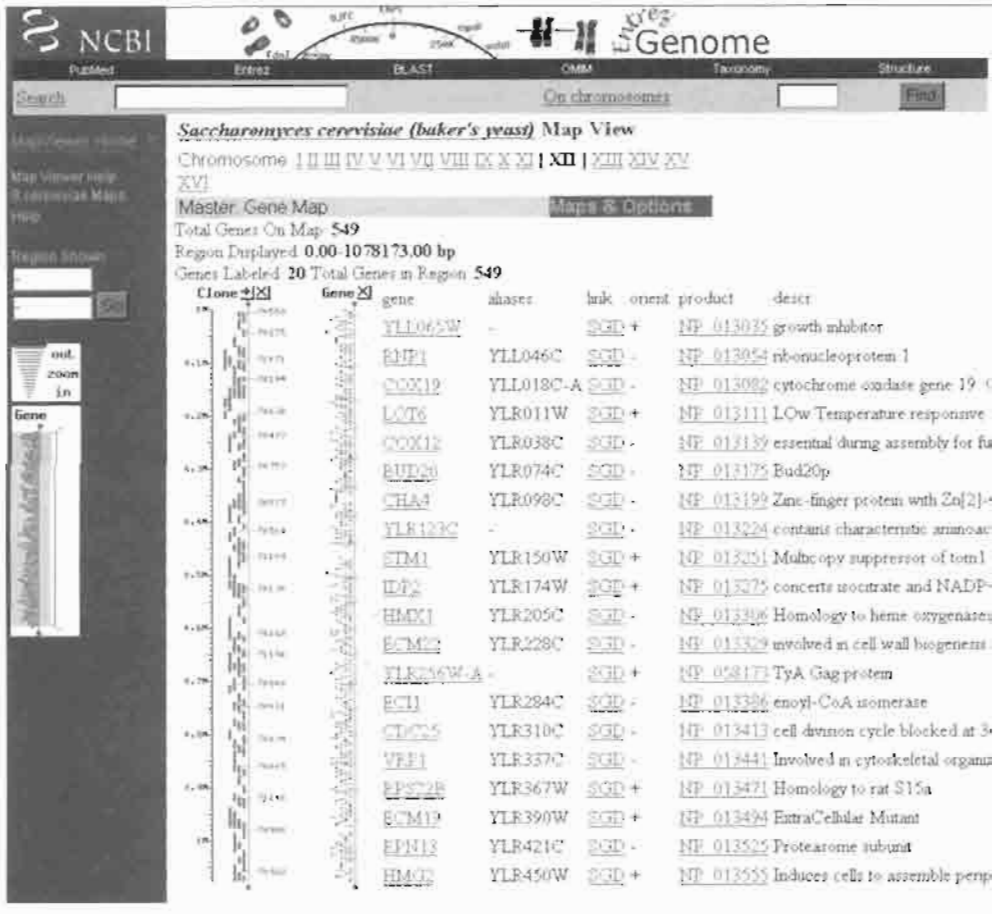


FIGURE 15.5. *You can access information on any* S. cerevisiae *chromosome through Entrez Genomes by selecting eukaryotes (or "prominent organisms") on the sidebar. The page for chromosome XII includes features such as COG, protein structure links, and TaxPlot that we have seen for bacteria and archaea.*

chromosomes. There are 534 ORFs of 100 or more codons on chromosome XII, with an average codon size of 485 codons.

Chromosome XII includes the largest gene in the *S. cerevisiae* genome, *YLR106c*. This gene encodes a protein with 4910 amino acids (MDN1p; accession Q12019) (Garbarino and Gibbons, 2002). Midasin, a human ortholog, is 5596 amino acids long (over 600 kD; RefSeq accession NP_055426).

For the nomenclature system used for *S. cerevisiae* genes and proteins, see Box 15.2.

Acquisition of New Genes: Duplication of *S. cerevisiae* Genome

As the genome sequence of *S. cerevisiae* was analyzed, it became apparent that there are many duplications of DNA sequence, involving both ORFs and larger genomic regions. In many cases, the gene order and orientation (top or bottom strand) is preserved between the duplicated regions. The duplications are both intrachromosomal and interchromosomal.

These changes in genetic material are fundamental in explaining the evolution of species in yeast or in any branch of life. We will see that in the human genome and a variety of other eukaryotic genomes, as many as 25% of the genes are duplicated (Chapters 16 and 17). From where can new, duplicate genes arise? Several mechanisms are outlined in Figure 15.8:

- genes can evolve by tandem repeat slippage during replication.
- New genes can arise by gene conversion. In this process, genes are transferred non-reciprocally from one genomic region to another. This occurs between repetitive regions of the human Y chromosome (Rozen et al., 2003).

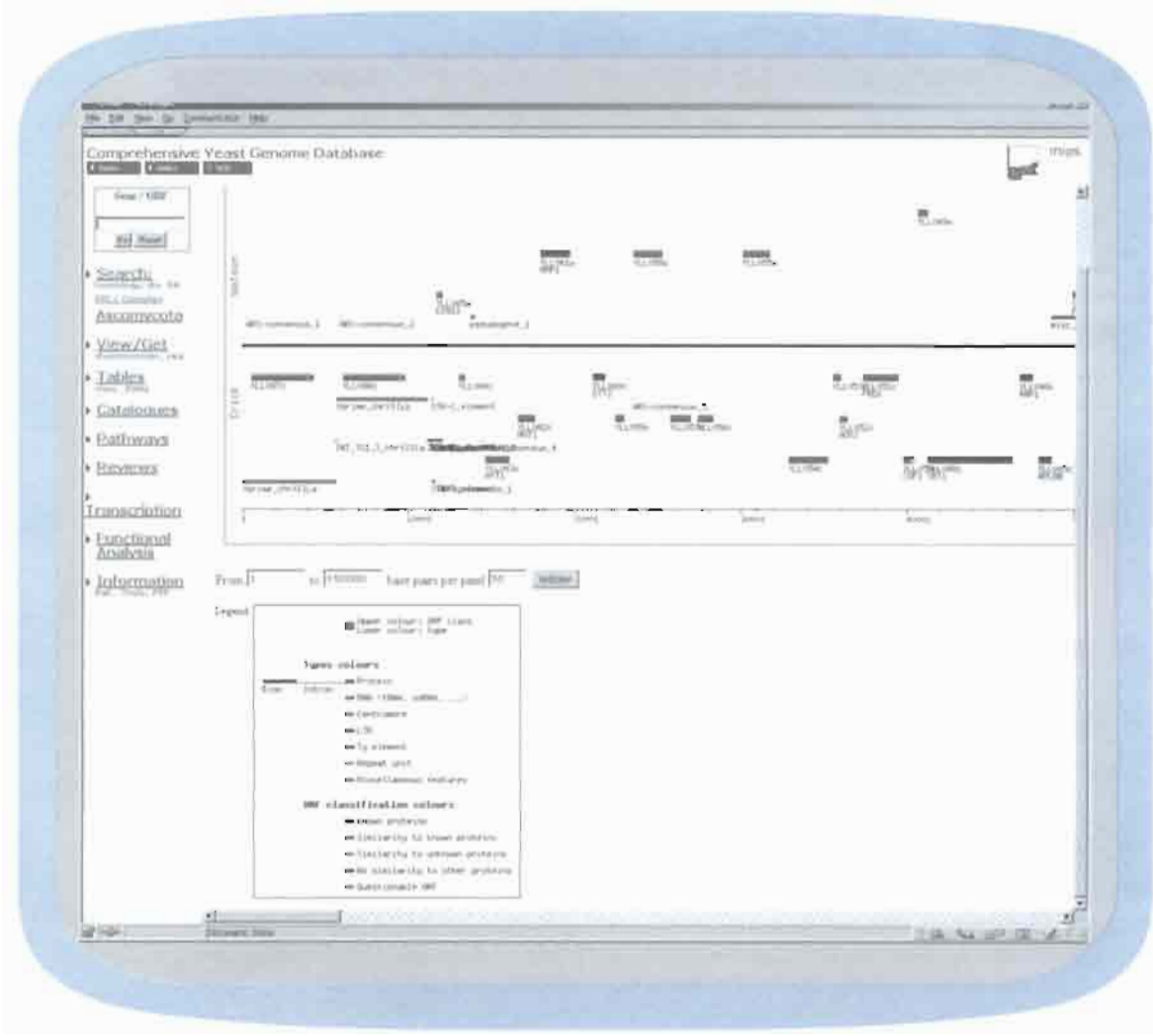


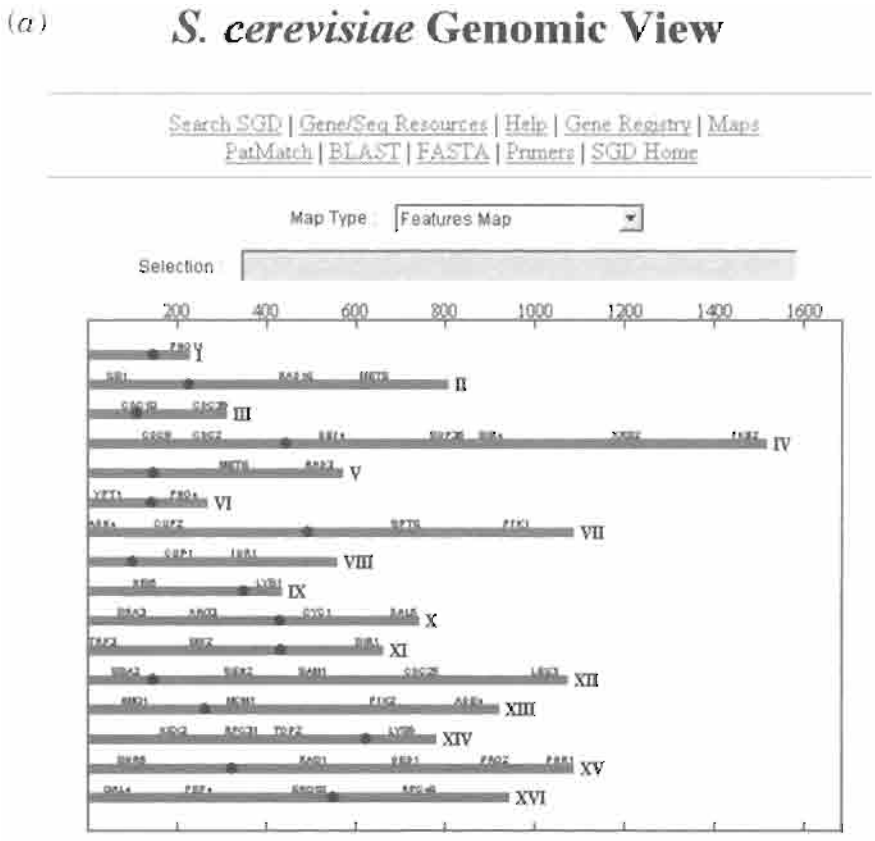
FIGURE 15.6. *MIPS offers the Comprehensive Yeast Genome Database. Each chromosome can be viewed on a map. The features are clickable for detailed information on each element, and the genes are annotated as described in Table 15.4.*

Tetraploidy is the presence of four haploid sets of chromosomes in the nucleus.

- genes can be introduced into a genome by lateral (horizontal) gene transfer.
- Segments of a genome can duplicate.
- An entire genome can duplicate, a process called polyploidy. In the case of *S. cerevisiae*, this is a tetraploidization.

In 1970, Susumu Ohno published *Evolution by Gene Duplication*. He proposed that vertebrate genomes evolved by two rounds of whole-genome duplication. These duplication events, according to this hypothesis, occurred early in vertebrate evolution and allowed the development of a variety of cellular functions. In his preface, Ohno (1970) wrote: "Had evolution been entirely dependent upon natural selection, from a bacterium only numerous forms of bacteria would have emerged. The creation of metazoans, vertebrates, and finally mammals from unicellular organisms would have been quite impossible, for such big leaps in evolution required the creation of new gene loci with previously nonexistent function. Only the cistron that became redundant was able to escape from the relentless pressure of natural

(a)



(b)

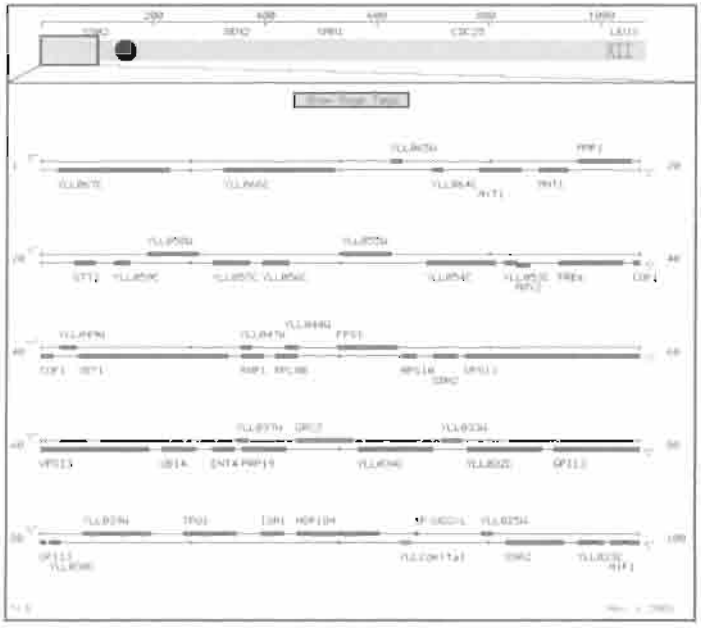


FIGURE 15.7. (a) *The SGD allows you to browse by chromsome.* (b) *Features on chromosome XII include protein-coding genes, RNAs (e.g., rRNA, tRNA, snoRNA), transposons, and SAGE tags.*

BOX 15-2
Gene Nomenclature in *Saccharomyces cerevisiae*

All ORFs that are ≥100 codons were assigned unique names consisting of three letters followed by a numeral and a subscript to describe its genomic position. For example, the gene name *YKL159c* refers to the ORF number 159 (from the centromere) on the left arm (*L*) of chromosome XI (*K*) of yeast (*Y*). The designations *c* or *w* ("Crick" or "Watson") reflect the orientation of the gene on the chromosome. Once a gene has been characterized and assigned some kind of function, the investigators may assign a new name that reflects the function, in this case *RCN1* for "regulator of calcineurin." Dominant alleles (typically the wild-type allele) are listed with three uppercase letters while recessive alleles (typically knockout mutations or loss of function alleles) are listed with three lowercase letters. The protein product of the gene is designated without italics and with only the first letter in uppercase and with "p" appended to designate protein. Many genes have multiple names (synonyms) because investigators have identified them in independent functional screens. Some examples of nomenclature are given in Table 15.5.

TABLE 15-5 Examples of Nomenclature for Yeast Genes

Wild-Type Allele	Protein Product	Mutant Alleles
CNA1	Cna1p	*cna1Δ*
RCN1	Rcn1p	*rcn1*, *rcn1::URA3*
YKL159c	Ykl159cp	*ykl159c*

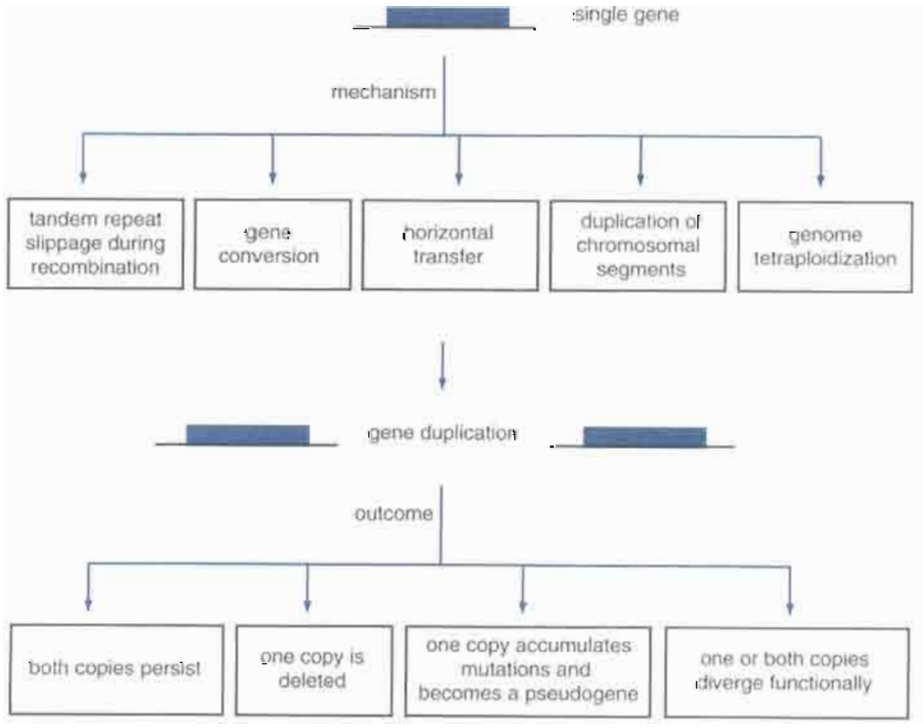


FIGURE 15.8. Gene duplication can occur by a variety of mechanisms. The duplicated copy may be localized on the same chromosome or on a different chromosome. There are several possible fates of a gene pair that arises by duplication: Copies may persist, one copy may be deleted, one copy may become nonfunctional (a pseudogene), or the two genes may acquire distinct functions. See Sankoff (2001).

selection. By escaping, it accumulated formerly forbidden mutations to emerge as a new gene locus."

Which mechanism of gene duplication might have occurred in *S. cerevisiae*? Wolfe and Shields (1997) provided support for Ohno's whole-genome duplication paradigm. They assessed the duplicated regions of the yeast genome by performing systematic blastp searches of all yeast proteins against each other and plotting the matches on dot matrices. Duplicate regions were observed as diagonal lines, such as the three duplicated regions seen in a comparison of proteins derived from chromosomes X and XI (Fig. 15.9). In the whole genome, they identified 55 duplicated regions and 376 pairs of homologous genes. In subsequent studies, they employed

Wolfe and Shields (1997) used Blastp rather than Blastn to study duplicated regions of chromosomes. This is because protein sequence data are more informative than DNA for the detection of distantly related sequences. See Chapter 3 (pages 41–42).

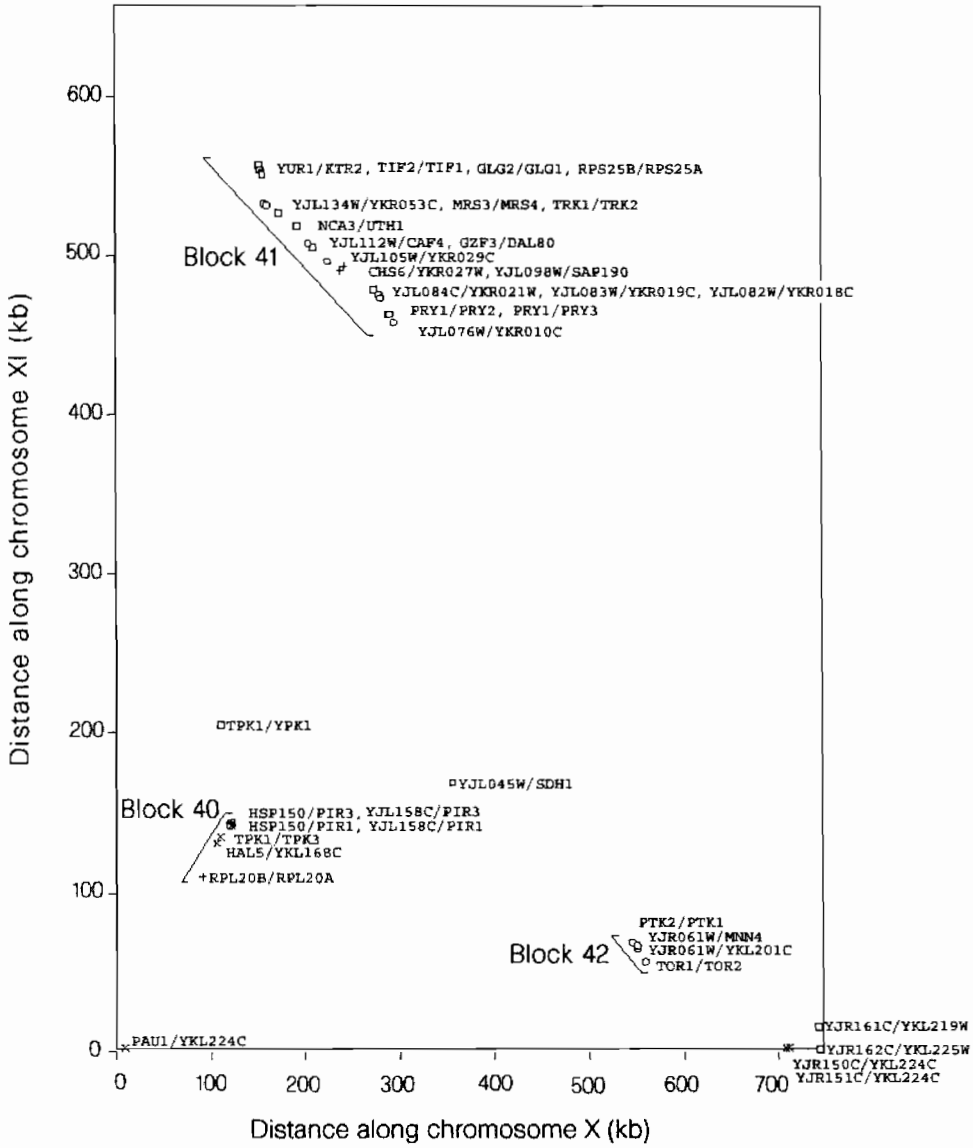


FIGURE 15.9. *Wolfe and Shields (1997) performed blastp searches of proteins from* S. cerevisiae *and found 55 blocks of duplicate regions. This provides strong evidence that the entire genome underwent an ancient duplication. This figure depicts the result of BLAST searches of proteins encoded by genes on chromosomes X and XI. Matches with scores >200 are shown, arranged in several blocks of genes. Symbols indicate the gene orientations: +, W (Watson strand orientation) on both chromosomes; x, C (Crick) orientation on both strands; squares and W indicate different orientations of the genes on the two chromosomes. Used with permission.*

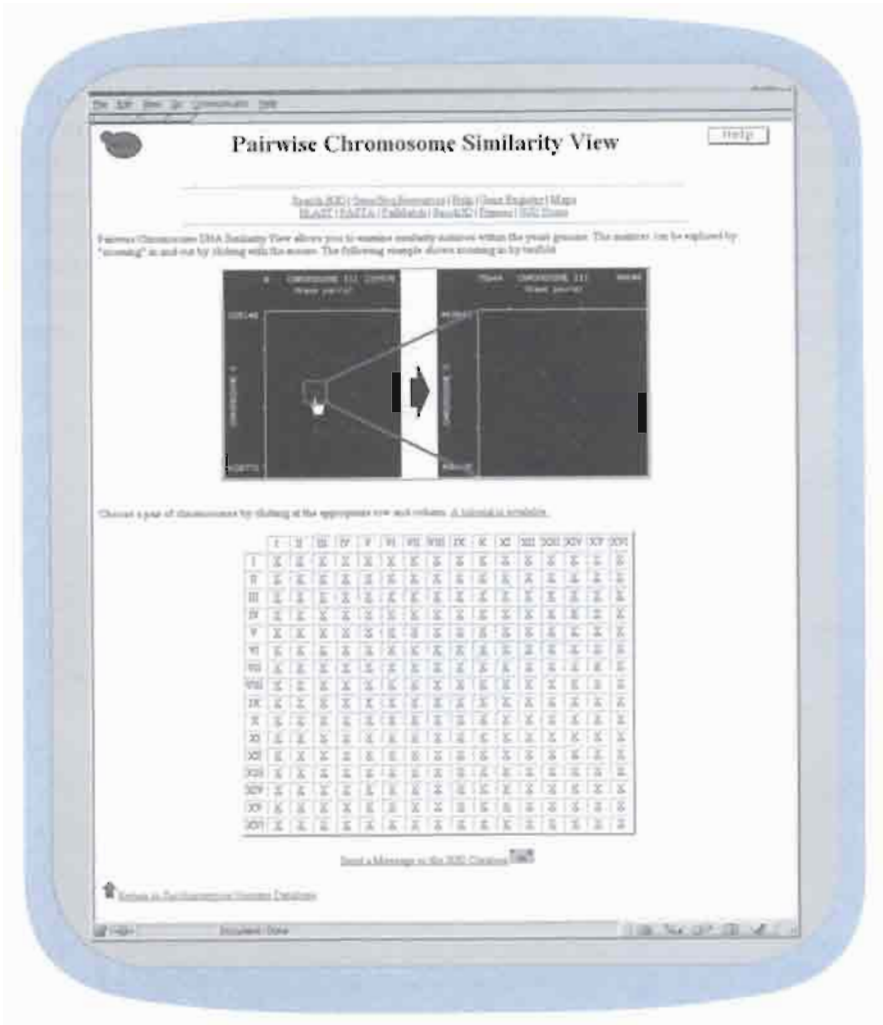


FIGURE 15.10. The SGD website includes a "sequence analysis and tools" section that includes a pairwise chromosome similarity viewer. This allows you to compare related regions on any two chromosomes, showing extended regions of duplication.

Kenneth Wolfe and colleagues (Trinity College, Dublin) have made available a website that allows pairwise comparisons of each *S. cerevisiae* chromosome with dot matrices. See ▸http://acer.gen.tcd.ie/~khwolfe/yeast/. You can also view chromosome comparisons at the SGD website (Fig. 15.10).

the more sensitive Smith–Waterman algorithm and identified a few additional regions of duplication (Seoighe and Wolfe, 1999). Based on these results, they proposed a single, ancient duplication of the *S. cerevisiae* genome (Wolfe and Shields, 1997). Subsequent to this duplication event, many duplicated genes were deleted. Other genes were rearranged by reciprocal translocation.

There are two main explanations for the presence of so many duplicated regions. There could have been whole-genome duplication (tetraploidy) followed by translocations, or alternatively there could have been a series of independent duplications. Wolfe and Shields (1997) favored the tetraploidy model for two reasons:

1. For 50 of the 55 duplicate regions, the orientation of the entire block was preserved with respect to the centromere. If each block was generated independently, a random orientation is expected.
2. Fifty-five successive, independent duplications of blocks would be expected to result in about seven triplicated regions, but only zero (or possibly one) such triplicated region was observed.

The model of genome duplication, with subsequent gene loss and rearrangements, is supported by a comparative analysis of 14 hemiascomycetes (fungi), including *S. cerevisiae* and 13 other species partially sequenced as part of the Génolevures project (Table 15.6).

TABLE 15-6 Génolevures Comparative Genomics Project

This project involves the sequencing of genomic DNA from 13 closely related fungi, all of the hemiascomycetes class. The site (▸http://cbi.labri.u-bordeaux.fr/Genolevures/index.php) is searchable by text queries and BLAST

Organism	Chromosome Number
Saccharomyces sensu stricto	
Saccharomyces bayanus var *uvarum*	16
Saccharomyces sensu lato	
Saccharomyces exiguus	14–16
Saccharomyces servazzii	7–12
Zygosaccharomyces rouxii	7
Saccharomyces kluyveri	8
Kluyveromyces	
Kluyveromyces thermotolerans	7
Kluyveromyces lactis	6
Kluyveromyces marxianus var *marxianus*	10
Distant species	
Pichia angusta	?
Debaromyces hansenii var *hansenii*	?
Pichia sorbitophila	7
Candida tropicalis	12
Yarrowia lipolytica	6

By comparing the gene order among these 14 organisms, Wong et al. (2002) mapped 70% of the *S. cerevisiae* genome to sister regions that have only minimal overlap. The 16 centromeres form 8 pairs, consistent with a model in which there was a whole-genome duplication. These analyses also suggest that several hemiascomycetes diverged from *S. cerevisiae* prior to its whole-genome duplication (*Kluyveromyces lactis*, *Zygosaccharomyces rouxii*, *Saccharomyces kluyveri*). Other hemiascomycetes such as *Saccharomyces bayanus* diverged after the hypothesized polyploidization event. Based on the evidence for polyploidy in several species, in combination with phylogenetic analyses, Wolfe and Shields (1997) estimated that the whole-genome duplication occurred about 100 MYA.

What is the fate of genes after duplication? The presence of extra copies of genes is usually deleterious to an organism. In the model of Wolfe and colleagues, the genome of an ancestral yeast doubled (from the diploid number of about 5000 to the tetraploid number of 10,000 genes) then lost the majority of its duplicated genes, yielding the present-day number of about 6200 ORFs. Overall, between 50 and 92% of duplicated genes are eventually lost (Wagner, 2001). For eukaryotes, the half-life of duplicated genes is only a few million years (Lynch and Conery, 2000) (see Chapter 16). There are four main possibilities (Fig. 15.8):

- Both copies can persist, maintaining the function of the original gene. In this scenario, there is a gene dosage effect because of the extra copy of the gene.
- One copy could be deleted. This appears to be the most common fate of duplicated genes. A rationale for this fate is that since the duplicated genes share identical functions initially, either one of them may be subject to loss-of-function mutations (Wagner, 2001).
- One copy can accumulate mutations and evolve into a pseudogene (a gene that does not encode a functional gene product). This represents a loss of gene function, although it occurs without the deletion of the duplicate copy.

The Génolevures project is described in a series of papers in *FEBS Letters* (December 22, 2000), volume 487, issue 1. The project website is ▸http://cbi.labri.fr/Genolevures/.

Saccharomyces cerevisiae can live under anaerobic conditions, while *K. lactis* cannot. It is possible that the *S. cerevisiae* genome duplication resulted in physiological changes that allowed this organism to acquire the new growth phenotype (Piskur, 2001).

In humans, an extra copy of chromosome 21 (i.e., trisomy 21) causes Down syndrome. Trisomies 13 and 18 are also sometimes compatible with life, but other autosomal trisomies are not. Duplications of even limited regions of the genome cause mental retardation and other diseases (see Chapter 18 on human disease). This highlights the deleterious nature of duplications at the level of individual organisms.

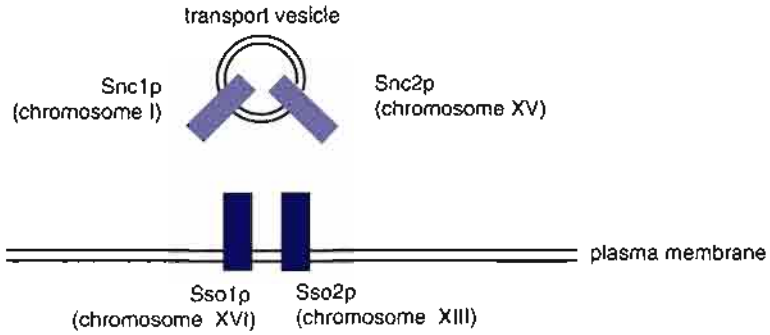


FIGURE 15.11. *The* S. cerevisiae *genes encoding a vesicle protein* (SNC1 *and* SNC2 *genes) and a target membrane docking protein* (SSO1 *and* SSO2). *At least one vesicle protein and one docking protein bind as part of a multimeric complex that mediates vesicle trafficking. The paralogous* SNC1/2 *and* SSO1/2 *genes may have arisen after a whole genome duplication (or alternatively after independent, segmental duplication events). The presence of two copies of each molecule could allow functional redundancy, so that if one copy is lost (e.g., through mutation), the organism could be viable. Alternatively, the duplicated genes could acquire distinct functions, such as conferring the specificity of the docking and fusion events of transport vesicles with the appropriate intracellular target membrane.*

- One or both copies of the gene could diverge functionally. According to this hypothesis, gene duplications (regardless of mechanism) could provide an organism with the raw material needed to expand its repertoire of functions. Furthermore, loss of either gene having overlapping functions might not be tolerated. Thus the functionally diverged genes would both be positively selected.

After a gene duplicates, why does one of the members of the newly formed gene pair often become inactivated? At first glance, it might seem highly advantageous to have two copies, because one may functionally diverge (driving the process of evolution to allow a cell to perform new functions), or one may be present in an extra copy in case the other undergoes mutation. However, gene duplication instead appears to be generally deleterious, leading to the loss of duplicated genes. The logic is that some mutations in a gene are *forbidden* rather than *tolerable* (these terms were used by Ohno [1970] in describing gene duplication). Forbidden mutations severely affect the function of a gene product, for instance by altering the properties of the active site of an enzyme. (A tolerable mutation causes a change that remains compatible with the function of the gene product.) Natural selection can eliminate forbidden mutations, because the individual is less fit to reproduce. After a gene duplicates, a deleterious mutation in one copy of a gene might now be tolerated because the second gene can assume its function. A second reason that duplicated genes may be deleterious is that in their presence the crossing over of homologous chromosomes during meiosis may be mismatched, causing unequal crossing-over.

We can consider the possible fates of duplicated genes with the specific example of genes encoding proteins that are essential for vesicle trafficking (Fig. 15.11). In yeast and all other eukaryotes, spherical intracellular vesicles transport various cargo to destinations within the cell. These vesicles traffic cargo to the appropriate target membrane through the binding of vesicle proteins (e.g., Snc1p in yeast or VAMP/synaptobrevin in mammals) to target membrane proteins (e.g., Sso1p in yeast or syntaxin in mammals) (Protopopov et al., 1993; Aalto et al., 1993). In *S. cerevisiae*, genome duplications presumably caused the appearance of two paralogous genes in each case: *SNC1* and *SNC2* as well as *SSO1* and *SSO2*. The *SNC1* and *SNC2* genes are on corresponding regions of chromosomes I and XV, while the *SSO1* and *SSO2* genes are on chromosomes XVI and XIII, respectively.

We encountered syntaxin in Figure 8.16.

FIGURE 15.12. *The SGD record for* SSO1 *shows that it is a nonessential gene (the organism is viable when the gene is deleted), but the double null mutant is inviable. SGD is accessed at* ▸*http://genome-www.stanford.edu/Saccharomyces/.*

What could the consequences of genome duplication have been? The two pairs of syntaxin-like and VAMP/synaptobrevin-like yeast proteins might have maintained the same function of the original proteins (before genome duplication). A search for *SSO1* at the SGD website shows that the gene is nonessential (the null mutant is viable), but the double knockout is lethal (see Fig. 15.12). Thus, it is likely that these paralogs offer functional redundancy for the organism; in the event a gene is lost (e.g., through mutation), the organism can survive because of the presence of the other gene. Similarly, the *SNC1* null mutant is viable, but the double knockout of *SNC1* and *SNC2* is deficient in secretion.

SGD can be searched at ▸http://genome-www.stanford.edu/Saccharomyces/.

As an alternative explanation of the duplication of these genes, it is possible that whole-genome duplication provided the new genetic materials with which the intracellular secretion machinery could be diversified. Syntaxin and VAMP/synaptobrevin proteins function at a variety of intracellular trafficking steps, and these gene families diversified throughout eukaryotic evolution (Dacks and Doolittle, 2002).

Andreas Wagner (2000) addressed the question of how *S. cerevisiae* protects itself against mutations by one of two mechanisms: (1) having genes with overlapping functions (such as paralogs that maintain related functions) or (2) through the interactions of nonhomologous genes in regulatory networks. He found that genes whose loss of function caused mild rather than severe effects on fitness did not tend to have closely related paralogs. This is consistent with a model in which gene duplication does not provide robustness against mutations.

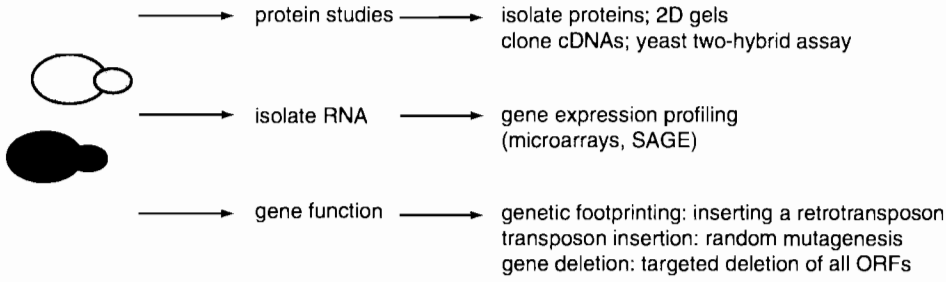


FIGURE 15.13. *Overview of functional genomics approaches in* S. cerevisiae. *See Ross-Macdonald (2000).*

FUNCTIONAL GENOMICS PROJECTS

Functional genomics refers to the assignment of function to genes based on genomewide screens and analyses. Once the protein-coding genes of an organism are identified, a major problem is to determine their function (see Chapter 8). Traditionally, function has been assessed one gene (or gene product) at a time by a combination of genetic, biochemical, and cell biological approaches. For example, thousands of proteins have been localized to cellular compartments in an effort to define their likely role(s), and homology searches have been used to define possible functional domains.

In *S. cerevisiae*, the functions of thousands of genes have been assessed in parallel using a variety of approaches at the level of genes, RNA transcripts, and proteins (Fig. 15.13). In Chapter 8, we described high-throughput approaches to defining protein–protein interactions in yeast, and in Chapters 6 and 7 we discussed SAGE and microarray-based approaches to gene expression. A variety of websites include expression data from experiments with *S. cerevisiae* (Table 15.7).

The Stanford Microarray Database is an important resource for yeast gene expression data (▸ http://www.dnachip.org/).

In this chapter we will describe several other high-throughput approaches to gene function through large-scale genetic screens. As a starting point, consider the *SSO1* gene entry at the SGD website (Fig. 15.12). The right sidebar of this web page includes pull-down menus that allow easy access to the results of functional genomics screens. These include protein–protein interaction data (Figs. 15.14 and 15.15), gene expression changes in a variety of conditions (Fig. 15.16), and many other analyses. There is additional structural information such as the relation of the genomic landscape to the corresponding genomic DNA of other fungi (Fig. 15.17).

We will next consider large-scale genetic approaches to function. Having the completed genome in hand, it is possible to create screens that comprehensively target the functional analysis of all known *S. cerevisiae* genes.

TABLE 15-7 Yeast Expression Data on World Wide Web

Subject	Reference	URL
Metabolic shift from fermentation to respiration	DeRisi et al., 1997	▸ http://cmgm.stanford.edu/pbrown/explore/index.html
Sporulation	Chu et al., 1998	▸ http://cmgm.stanford.edu/pbrown/sporulation/
Cell cycle	Spellman et al., 1998	▸ http://genome-www.stanford.edu/cellcycle/
Transcription	Holstege et al., 1998	▸ http://web.wi.mit.edu/young/expression/
Response to alkylating agent	Jelinsky and Samson, 1999	▸ http://www.hsph.harvard.edu/geneexpression/
Multidrug resistance	DeRisi et al., 2000	▸ http://www.biologie.ens.fr/fr/genetiqu/puces/publications/pdr1-3_3-7/
Microarray global viewer	Marc et al., 2001	▸ http://www.transcriptome.ens.fr/ymgv/

Source: From MIPS database (▸ http://mips.gsf.de/proj/yeast/CYGD/db/index.html, November 2002).

FIGURE 15.14. *The SGD offers links from a typical gene entry* (SSO1) *to known interactions based on the yeast two-hybrid assay (see Chapter 8). Sso1p binds to other proteins in heteromultimeric complexes that mediate vesicle docking and fusion. All the* genes *depicted in this figure have homologs in mammals that are involved in the docking and fusion of synaptic vesicles that release neurotransmitters. This highlights the usefulness of* S. cerevisiae *as a model organism to study biological processes in eukaryotes.*

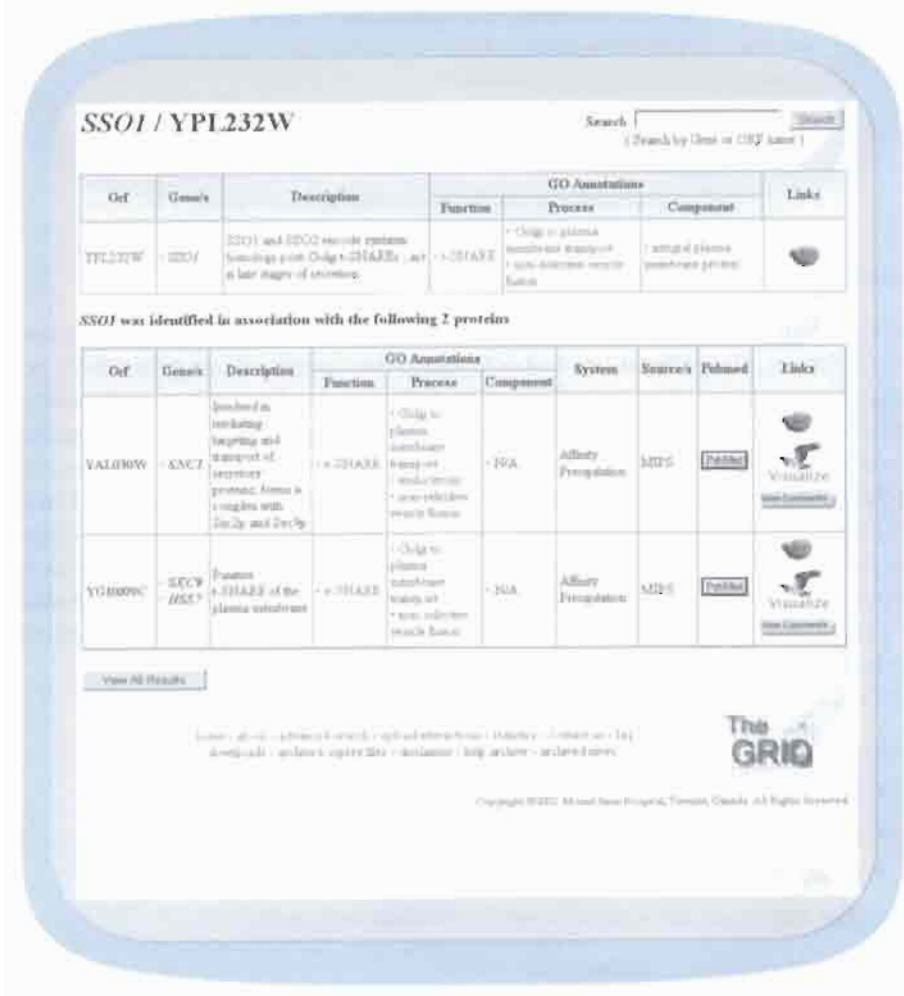


FIGURE 15.15. *The SGD entry for* SSO1 *links to the General Repository for Interaction Datasets (GRID) database which describes proteins that interact with Sso1p. The GRID URL is* ▸ *http://biodata.mshri.on.ca/grid/servlet/Index.*

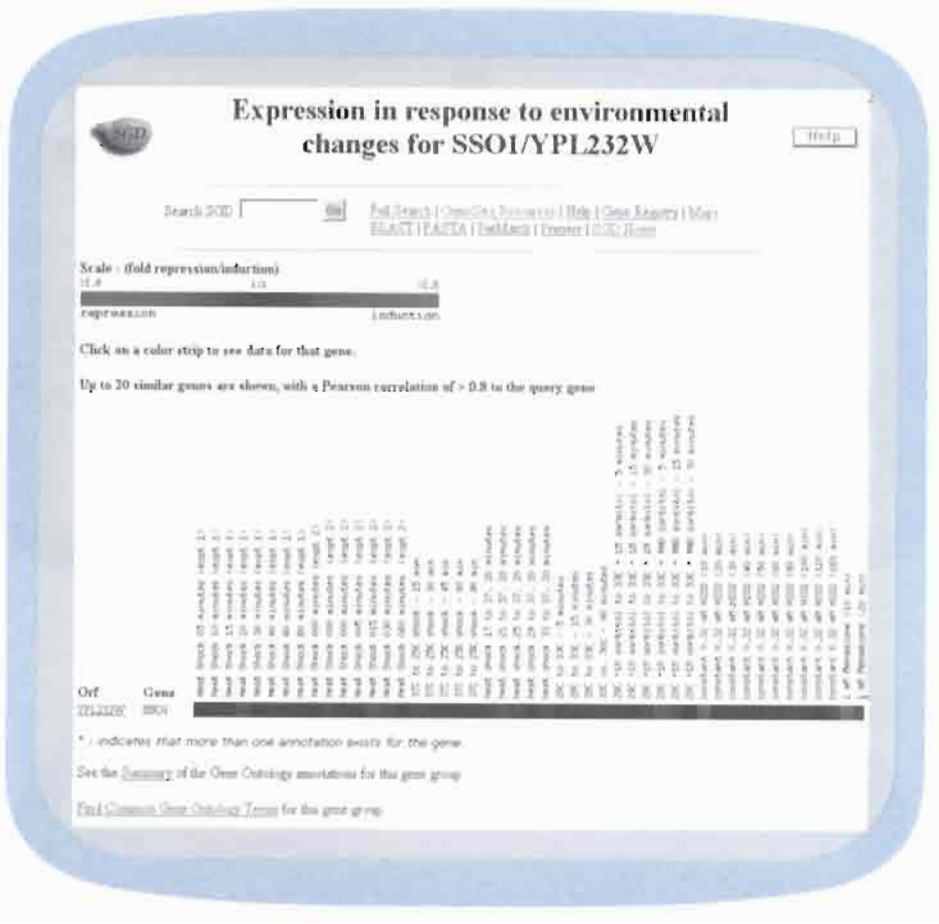


FIGURE 15.16. The SGD report for a typical gene (SSO1) *shows its changes in expression in response to many dozens of environmental changes.*

Genetic Footprinting Using Transposons

The Ty1 element is a yeast retrotransposon that inserts randomly into the genome. Patrick Brown, David Botstein, and colleagues developed a strategy in which populations of yeast are grown under several different conditions (e.g., rich medium versus minimal medium) and subjected to Ty1 transposon-mediated mutagenesis (Smith et al., 1995, 1996) (Fig. 15.18). Following the insertion, the polymerase chain reaction (PCR) is performed using primers that are specific to the gene and to the Ty1 element. This results in a series of DNA products of various molecular weights. The premise of the approach is that an individual gene (e.g., *SSO1*) might be important for growth under certain conditions. There will be a loss of PCR products (a "genetic footprint") that indicates the importance of that gene for a particular condition.

This approach has several advantages:

- Any gene of interest can be assayed or genes can be selected randomly.
- Multiple mutations can be assayed for any given gene.
- It is possible to perform phenotypic analyses in parallel in a population.
- Many different phenotypes can be selected for analysis.
- The approach can succeed even for overlapping genes.

There are also several disadvantages:

- Mutant strains are not recovered.

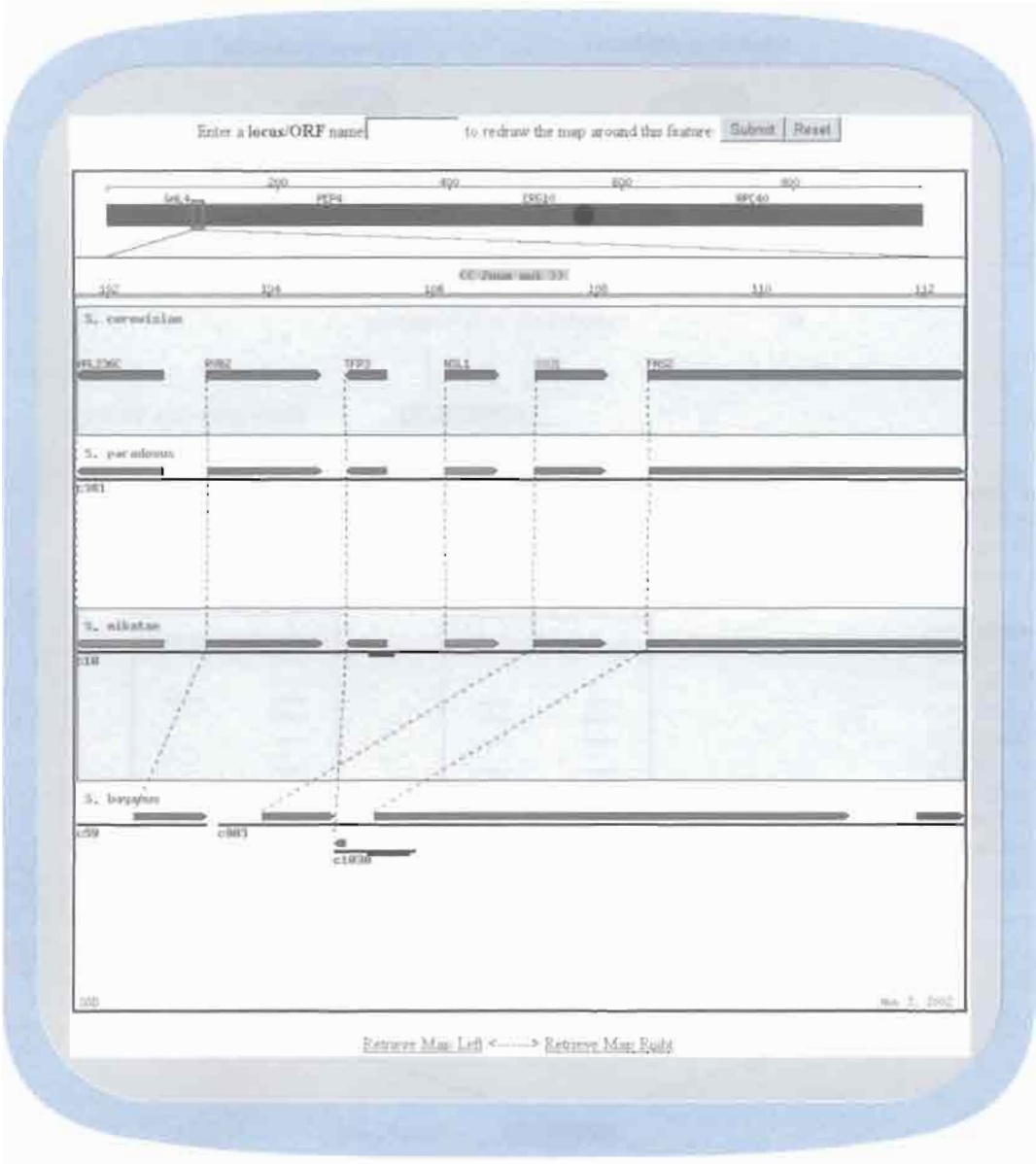
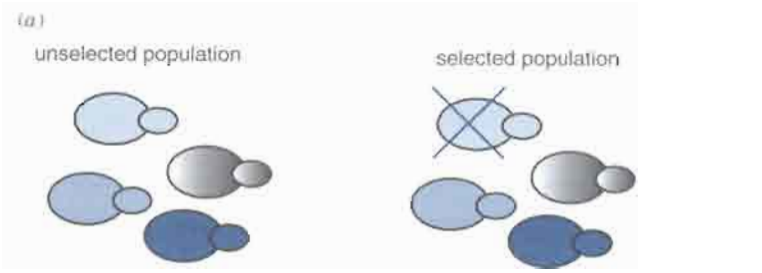


FIGURE 15.17. The synteny viewer at SGD shows the region of genomic DNA in chromosome XVI (where SSO1 is localized) and the corresponding (conserved syntenic) genomic DNA regions of related fungi.

- Multiple mutations (alleles) are generated, but they are all insertions (rather than knock-outs or other types of mutation).
- The approach is labor intensive and entails a gene-by-gene analysis.
- The role of duplicated genes with overlapping functions may be missed.

Harnessing Exogenous Transposons

Another mutagenesis approach involves the random insertion of tags into genes using bacterial or yeast transposons (Ross-Macdonald et al., 1999) (Fig. 15.19). A mini-transposon derived from a bacterial transposon Tn3 contains a *lacZ* reporter gene lacking an initiator methionine or upstream promoter sequence. When randomly inserted into a protein-coding gene, it is expected to be translated in-frame in one



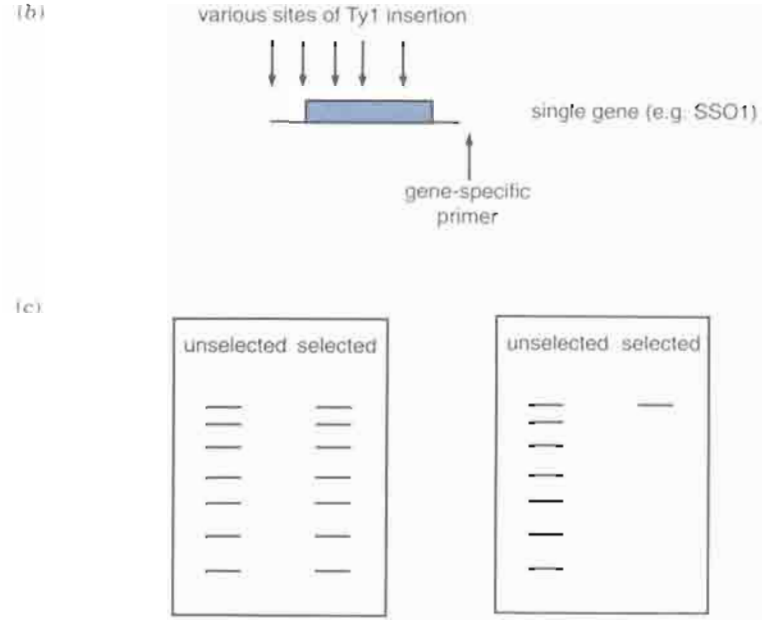
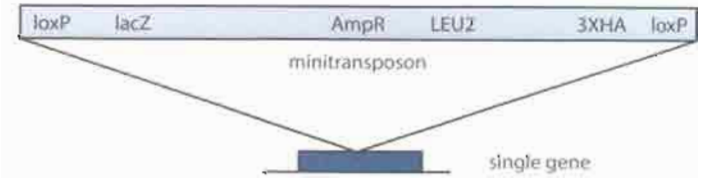


FIGURE 15.18. Genetic footprinting. (a) *A population of yeast is selected, e.g., by changing the medium or adding a drug. Some genes will be unaffected by the selection process.* (b) *Random insertion of a transposon allows gene-specific PCR to be performed and* (c) *subsequent visualization of DNA products electrophoresed on a gel. Some genes will be unaffected by the selection process (panel at left). Other genes, tagged by the transposition, will be associated with a reduction in fitness. Less PCR product will be observed [in* (c)*], thus identifying this gene as necessary for survival of yeast in that selection condition.*



mutagenize library (yeast genome)

generate lacZ-fusion protein strains
(study expression patterns and disruption phenotypes)

generate hemagglutinin-tagged strains
(study localization)

FIGURE 15.19. Transposon tagging and gene disruption to assess gene function in yeast. (Adapted from Ross-Macdonald et al., 1999.)

out of six cases. When this happens, the yeast will produce β-galactosidase, allowing the insertion event to be detected. The construct includes *loxP* sites that allow a recombination event in which the gene is tagged with only a short amount of DNA encoding three copies of a hemagglutinin (HA) epitope tag.

We described the HA tag in Chapter 8. An HA-tagged protein can be localized within a cell using an antibody specific to HA.

This minitransposon construct allows a genomewide analysis of disruption phenotypes, gene expression studies, and protein localization. Ross-Macdonald et al. (1999) generated 11,000 yeast strains in which they characterized disruption phenotypes under 20 different growth conditions. These studies resulted in the identification of 300 previously nonannotated ORFs. Data from this study were deposited in the TRIPLES database (Kumar et al., 2002b). An example of a search result from this database is shown for *SSO2* (Fig. 15.20).

The TRIPLES database is available at ▸http://ygac.med.yale.edu/triples/triples.htm. TRIPLES stands for Transposon-Insertion Phenotypes, Localization and Expression in *Saccharomyces*. As of 1999, the database has data on over 92,000 plasmid preps, 11,000 blue yeast strains (positive for β-galactosidase), 6300 sequenced yeast strains, 1900 known ORFs that are affected (i.e., with an insertion inside the gene or within 200 bp), and 1346 in-frame fusions to known ORFs.

This large-scale approach generated a variety of information:

- Surprisingly, 480 expressed insertions were fused to an ORF but in the wrong reading frame. This suggests that frame shifting may be a very common gene expression mechanism.

(*a*)

Query Results Help

The search conditions are: Gene Name: sso2
The output results are sorted by: Clone ID

Clone ID	Gene or ORF (plus synonyms)	mTn insertion	Insertion point	Total protein length
TN7-113D8	SSO2/YM8010.13/YMR183C	in-frame	94	295
TN7-95G9	SSO2/YM8010.13/YMR183C	out-of-frame	157	295
V195A4	SSO2/YM8010.13/YMR183C	in-frame	272	295
V10G6	SSO2/YM8010.13/YMR183C	out-of-frame	-116	0

- **Please click on the highlighted clone ID to obtain a composite report of phenotypic, expression, and protein localization data for a given clone.**
- **Please access our on-line help for a more complete description of the data presented above.**

(*b*)

V195A4 Request reagents

Genbank accession	SGD Link	YPD Link	Genecensus Link (Gerstein's Lab)
none	YMR183C	YMR183C	YMR183C

Potential ORFs Disrupted by mTn Insertion Help

Gene or ORF (plus synonyms)	mTn insertion/ORF features	Chr.	Chr. Insertion Coord.	Insertion point (codon #)	Total ORF length (amino acids)	Repeated Gene	LacZ Orientation
SSO2/YM8010.13/YMR183C	In frame	XIII	626998	272	295	may be	sense

Gene Expression Data Help

Growth condition	LacZ expression level	Background strain
vegetative	intense blue	Y800
sporulation	intense blue	Y800

Subcellular Localization Help

Localization	Cell stage	% of cells staining	Stain intensity	Trials	Comments
cyto. (patches) / Endo. Reticulum	all	35	2	2	patchy staining of the cytoplasm and ER

FIGURE 15.20. *(a) An example of a search result of the TRIPLES database for* SSO2, *(b) showing a mutant with a transposon insertion and a resultant phenotype.*

- There were 328 in-frame insertions to nonannotated ORFs (from 50 to 247 codons long). Thus, this method is useful to identify novel protein-coding genes.
- Fifty-two percent of these previously nonannotated ORFs are antisense to a known ORF, and 15% overlap a known ORF in a different frame. Thirty-three percent were intergenic. These findings are consistent with the hypothesis that many small genes remain undiscovered.

This approach offers a variety of useful features:

- It includes data on expression levels and protein localization.
- New genes can be discovered, as described above.
- The analysis works for overlapping genes.
- Mutant strains can be recovered and banked. They are made available to the scientific community through the TRIPLES database.

Problems are similar to those involved in genetic footprinting and include the requirement for transposon site specificity, the missing of information on genes with duplicated functions, and the labor-intensive nature of the project.

Genomewide Deletions with Molecular Barcodes

A consortium of researchers achieved the remarkable goal of creating yeast strains representing the targeted deletion of virtually every known gene (Winzeler et al., 1999; Giaever et al., 2002). The goals of this project are as follows:

- To create a yeast knock-out collection in which all of the $\approx$ 6000 ORFs in the *S. cerevisiae* genome are disrupted
- To provide all nonessential genes (85% of the total) in several useful forms: diploids heterozygous for each yeast knock-out, diploids homozygous for each yeast knock-out, a mating type, and α-mating-type strains
- To provide all essential genes (15% of the total) as diploids heterozygous for each yeast knock-out

Budding yeasts have two mating types: a, and α. Haploid a and α cells can mate with each other to form diploid a/α cells. Both haploid and diploid phases of the life cycle grow mitotically.

The strategy employed for this project is gene replacement by PCR, relying on the high rate of homologous recombination that occurs in yeast (Fig. 15.21). A short region of DNA (about 50 bp), corresponding to the upstream and downstream portions of each gene, is placed on the end of a selectable marker gene. Additionally, two "molecular barcodes"—an UPTAG and a DOWNTAG—are unique 20-bp oligonucleotide sequences added to each deletion strain. This feature allows thousands of deletion strains to be pooled and assayed in parallel in a variety of growth conditions. Giaever et al. (2002) described the genes that are necessary for optimal growth under six conditions: high salt, sorbitol, galactose, pH 8, minimal medium, and treatment with the antifungal drug nystatin. Among their findings:

- About 19% of the yeast genes (1105) were essential for growth on rich glucose medium. Only about half of these genes were previously known to be essential.
- Nonessential ORFs are more likely to encode a new protein.

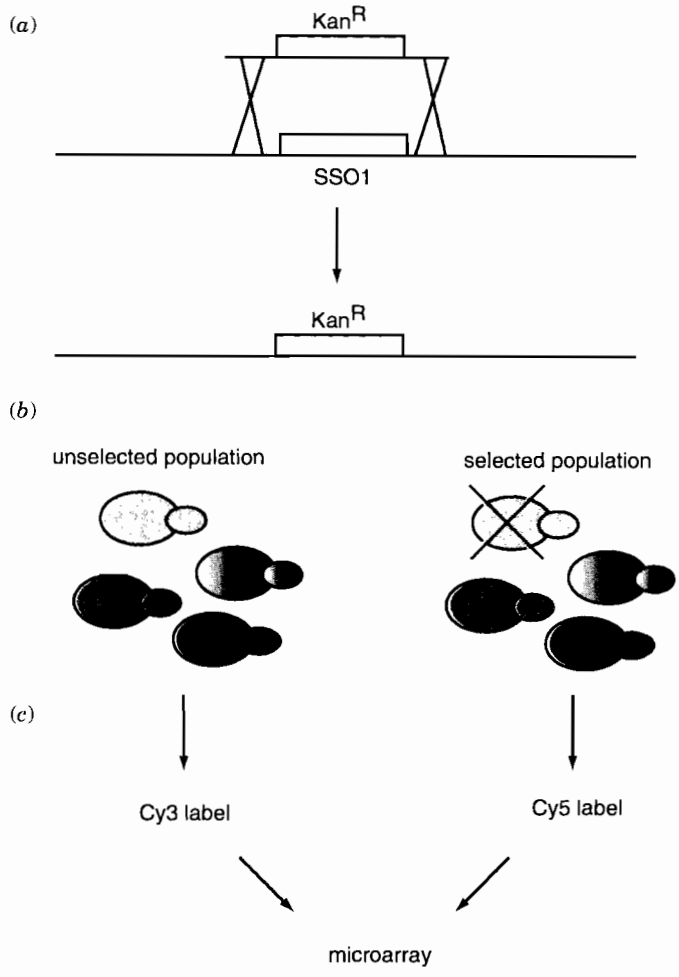


FIGURE 15.21. *Targeted deletion of virtually all* S. cerevisiae *genes.* (a) *The strategy is to use gene replacement by homologous recombination. Each gene (e.g.,* SSO1*) is deleted and replaced by a* KanR *gene, with unique UPTAG and DOWNTAG primer sequences located on either end.* (b) *A variety of selection conditions can be used.* (c) *Genomic DNA is isolated from each condition, labeled with Cy3 or Cy5, and hybridized to a microarray. In this way, genes functionally involved in each growth condition can be identified.*

- Essential genes are more likely to have homologs in other organisms.
- Few of the essential genes are duplicated within the yeast genome (8.5% of the nonessential genes have paralogs, while only 1% of the essential genes have paralogs). This supports the hypothesis that duplicated genes have important redundant functions (see pages 517–518 above).

The systematic deletion method offers a number of important advantages:

- All known genes in the *S. cerevisiae* genome are assayed.
- Each mutation is of a defined, uniform structure
- Mutations are guaranteed to be null.
- Mutant knock-out strains are recovered, banked, and made available to the scientific community.
- Studies of multigene families are facilitated.
- Parallel phenotypic analyses are possible, and many different phenotypes can be assayed.
- Once the strains have been generated, the labor requirement is low when a new phenotype is assessed.

This method also has limitations:

- The labor investment to generate these knock-outs was very large.
- For each gene, only null alleles were generated for study. (Additional alleles may be available from other studies.)
- No new genes are discovered with this approach, in contrast to random transposon insertion approaches.
- All nonannotated ORFs are missed. In particular, short ORFs may not be annotated (see above).
- Deletions in overlapping genes may be difficult to interpret.

Analysis of Fungal Genomes

In addition to *S. cerevisiae*, the genomes of many other fungi are now being sequenced (Table 15.8). We will discuss some of these projects in alphabetical order. The second fungal genome to be completely sequenced was *S. pombe*, followed by *Neurospora crassa*.

TABLE 15-8 Fungal Genome-Sequencing Projects

Organism	Size (Mb)	Chromosome Number	Source:
Aspergillus fumigatus	30	8	▸http://www.sanger.ac.uk/Projects/A_fumigatus/
			▸http://bioinf.man.ac.uk/
			▸http://www.pasteur.fr/externe
			▸http://www.tigr.org/
Aspergillus nidulans	28.5	8	▸http://www.cereon.com/
Aspergillus parasiticus	NA	NA	▸http://www.genome.ou.edu/fungal.html
Candida albicans	16	8	▸http://sequence-www.stanford.edu/group/candida/
Cryptococcus neoformans	21	NA	▸http://rcweb.bcgsc.bc.ca/cgi-bin/cryptococcus/cn.pl
			▸http://www.nagasaki-u.ac.jp/english/
			▸http://www-sequence.stanford.edu/group/C.neoformans/
			▸http://www.genome.ou.edu/cneo.html
Fusarium sporotrichioides	NA	NA	▸http://www.genome.ou.edu/fsporo.html
Magnaporthe grisea	40	7	▸http://www.wi.mit.edu/
			▸http://www.fungalgenomics.ncsu.edu/index.htm
Neurospora crassa	43	7	▸http://www-genome.wi.mit.edu/
			▸http://www.unm.edu/~ngp/
			▸http://www.genome.ou.edu/fungal.html
Phanerochaete chrysosporium	30	10	▸http://www.jgi.doe.gov/programs/whiterot.htm
Saccharomyces cerevisiae	12	16	▸http://genome-www.stanford.edu/Saccharomyces/
Schizosaccharomyces pombe	13.8	3	*Nature*, vol. 415, pp. 871–880, 2002
Ustilago maydis	20	NA	LION Bioscience and Bayer

Source: From NCBI.
Abbreviation: NA, not available.

Aspergillus fumigatus

Aspergillus fumigatus is the most common mold that causes infection worldwide. It is classfied as a filamentous fungus. Its genome is being sequenced by collaboration between The Institute for Genomic Research, The Wellcome Trust Sanger Institute, and the Pasteur Institut.

A TIGR website provides an overview of *A. fumigatus* and the current status of the genome sequencing (▸http://www.tigr.org/tdb/e2k1/afu1/). The *Aspergillus* website (▸http://www.aspergillus.man.ac.uk/) provides detailed information.

Candida albicans

Candida albicans is a diploid sexual fungus that frequently causes opportunistic infections in humans. The genome size is approximately 16 Mb. The chromosomal arrangement is unusual: The genome has eight chromosome pairs, seven of which are constant and one of which is variable (ranging from about 3 to 4 Mb).

The *C. albicans* genome is being sequenced at Stanford University (▸http://sequence-www.stanford.edu/group/candida/index.html). Further information on this fungus is available from ▸http://alces.med.umn.edu/Candida.html.

Atypical Fungus: Microsporidial Parasite *Encephalitozoon cuniculi*

Microsporidia are single-celled eukaryotes that lack mitochondria and peroxisomes. These organisms infect animals (including humans) as obligate intracellular parasites. The complete genome of the microsporidium *E. cuniculi* was determined by several research groups in France (Katinka et al., 2001). The genome is highly compacted, having about 2000 protein-coding genes in 2.9 Mb. Thus, analogous to parasitic bacteria (Chapter 14), these pathogens have undergone a reduction in genome size. Phylogenetic analyses using several *E. cuniculi* proteins suggest that these parasites are atypical fungi that once possessed but subsequently lost their mitochondria (Fig. 15.22) (Katinka et al., 2001).

Neurospora crassa

The orange bread mold *Neurospora* has served as a beautiful and simple model organism for genetic and biochemical studies since Beadle and Tatum used it to

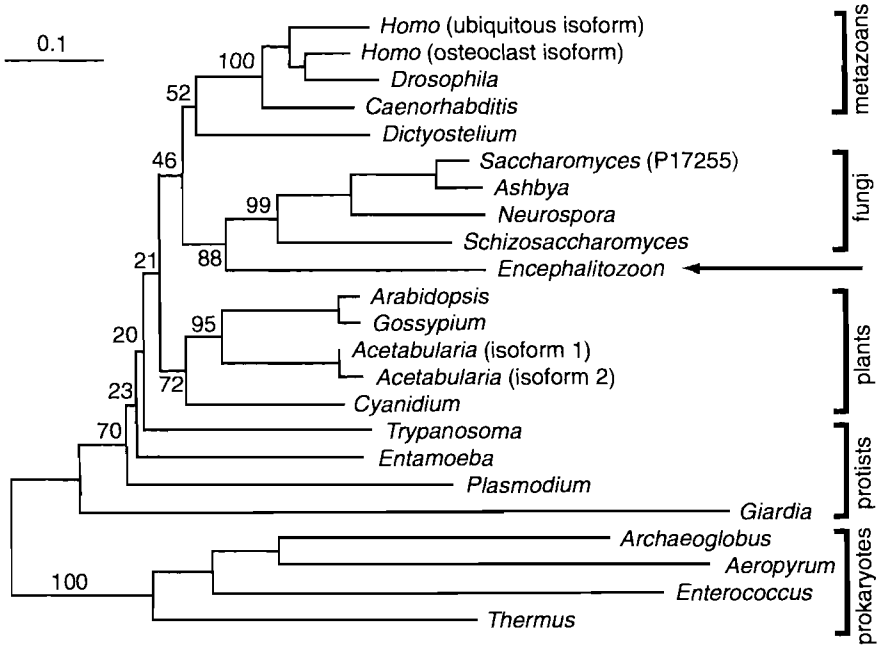


FIGURE 15.22. *Phylogenetic analysis of vacuolar ATPase subunit A from animals, plants, fungi, protists, and prokaryotes supports a fungal origin for the microsporidial parasite* Encephalitozoon cuniculi *(arrow). This tree was generated using a neighbor-joining method by Katinka et al. (2001). Values are bootstrap percentages (see Chapter 11). Used with permission.*

Neurospora crassa genome database websites are available at the University of New Mexico (▸ http://www.unm.edu/~ngp/), at the Whitehead Institute (▸ http://www-genome.wi.mit.edu/annotation/fungi/neurospora/), and at MIPS (▸ http://www.mips.biochem.mpg.de/proj/neurospora/).

George Beadle and Edward Tatum shared a Nobel Prize in 1958 (with Joshua Lederberg) "for their discovery that genes act by regulating definite chemical events" (▸ http://www.nobel.se/medicine/laureates/1958/). They irradiated *N. crassa* with X rays to study gene function.

establish the one gene–one enzyme model in the 1940s. *Neurospora* is the best characterized of the filamentous fungi, a group of organisms critically important to agriculture, medicine, and the environment (Perkins and Davis, 2000). The developmental complexity of *Neurospora* contrasts with other unicellular yeasts (Casselton and Zolan, 2002). *Neurospora* is widespread in nature and thus, like the fly *Drosophila*, it is exceptionally suited as a subject for population studies.

Like *S. cerevisiae*, *Neurospora* is an ascomycete and thus shares the advantage of this group of organisms in yielding complete tetrads for genetic analyses in the laboratory. However, it is more similar to animals than yeasts in many important ways. For example, unlike yeast but like mammals, it contains complex I in its respiratory chain, it has a clearly discernable circadian rhythm, and it methylates DNA to control gene expression. The six decades of intensive studies on the genetics, biochemistry, and cell biology of *Neurospora* establish this organism as a gold mine of biological knowledge.

Galagan et al. (2003) reported the complete genome sequence of *Neurospora*. They sequenced about 39 Mb of DNA on 7 chromosomes, and identified 10,082 protein-coding genes (9,200 longer than 100 amino acids). Of these proteins, 41% have no similarity to known sequences, and 57% do not have identifiable orthologs in *S. cerevisiae* or *S. pombe*.

The *Neurospora* genome has only 10% repetitive DNA, including ~185 copies of rDNA genes (Krumlauf and Marzluf, 1980). Other repeated DNA is dispersed and tends to be short and/or diverged, presumably because of the phenomenon of "RIP" (repeat-induced point mutation). RIP is a mechanism by which the genome is scanned for duplicated (repeated) sequences in haploid nuclei of special premeiotic cells. The RIP machinery efficiently finds them, and then litters them with numerous GC-to-AT mutations (Selker, 1990). Apparently RIP serves as a genome defense system for *Neurospora*, inactivating transposons and resisting genome expansion (Kinsey et al., 1994). Galagan et al. (2003) found relatively few *Neurospora* genes that are in multigene families, and a mere eight pairs of duplicated genes that encode proteins >100 amino acids. Also, 81% of the repetitive DNA sequences were mutated by RIP. Thus, RIP has suppressed the creation of new genes through duplication in *Neurospora* (Galagan et al., 2003; Perkins et al., 2001).

The *P. chrysosporium* genome-sequencing project is being undertaken by the U.S. Department of Energy (▸ http://www.jgi.doe.gov/programs/whiterot.htm). We introduced a white rot BLAST server in Chapter 5.

First Basidiomycete: *Phanerochaete chrysosporium*

Phanerochaete chrysosporium is the first fungus of the phylum Basidiomycota to have its genome completely sequenced. This is a white rot fungus that degrades many biomaterials, including pollutants. The genome consists of about 30 Mb of DNA arranged in 10 chromosomes.

For extensive information on *S. pombe* genome sequence analysis, see The Wellcome Trust Sanger Institute website (▸ http://www.sanger.ac.uk/Projects/S_pombe/). TIGR offers the *Schizosaccharomyces* Gene Index (▸ http://www.tigr.org/tdb/spgi/).

Fission Yeast *Schizosaccharomyces pombe*

The fission yeast *S. pombe* has a genome size of 13.8 Mb. The complete sequencing of this genome was reported by a large European consortium (Wood et al., 2002). The genome is divided into three chromosomes (Table 15.9).

Notably, there are 4940 predicted protein-coding genes (including 11 mitochondrial genes) and 33 pseudogenes. This is substantially fewer genes than is found in *S. cerevisiae* and represents the smallest number of protein-coding genes observed

TABLE 15-9 Features of *S. pombe* Genome

Chromosome Number	Length (Mb)	Number of Genes	Mean Gene Length (bp)	Coding (%)
1	5.599	2255	1446	58.6
2	4.398	1790	1411	57.5
3	2.466	884	1407	54.5
Whole genome	12.462	4929	1426	57.5

Source: From Wood et al. (2002).

for any eukaryote. In fact, several bacterial genomes encode more proteins: *Mesorhizobium loti* (6752 predicted genes) and *Streptomyces coelicolor* (7825 predicted genes).

The gene density in *S. pombe* is about one gene per 2400 bp, which is slightly less dense than is seen for *S. cerevisiae*. The intergenic regions are longer, and about 4730 introns were predicted. In *S. cerevisiae*, only 4% of the genes have introns.

Schizosaccharomyces pombe and *S. cerevisiae* diverged between 330 and 420 MYA. Some gene and protein sequences are equally divergent between these two fungi as they are between fungi and their human orthologs. To see this, you can use the TaxPlot tool on the NCBI Entrez genomes website (Fig. 15.23). In this way, it is easy to identify fungal genes that are most closely related to humans. Comparative analyses are likely to elucidate the genetic basis for differences in the biology of these fungi, such as the propensity of *S. pombe* to divide by binary fission and the relatively fewer number of transposable elements in *S. pombe*.

Leland Hartwell, Timothy Hunt, and Sir Paul Nurse won the Nobel Prize in Physiology or Medicine in 2001 for their work on cell cycle control. Nurse's studies employed *S. pombe*, while Hartwell studied *S. cerevisiae* and Hunt studied sea urchins and other organisms. See ▸http://www.nobel.se/medicine/laureates/2001/.

PERSPECTIVE

The budding yeast *S. cerevisiae* is one of the most significant organisms in biology for several reasons:

- It represents the first eukaryotic genome to have been sequenced. It was selected because of its compact genome size and structure.
- As a single-celled eukaryotic organism, its biology is simple relative to humans and other metazoans.
- The biology community has acquired a deep knowledge of yeast genetics and has collected a variety of molecular tools that are useful to elucidate the function of yeast genes. Functional genomics approaches based on genomewide analysis of gene function are now being implemented. For example, each of its >6000 genes has been knocked out and tagged with molecular barcodes, allowing massive, parallel studies of gene function.

Many additional fungal genomes are now being sequenced. In each branch of biology, we are learning that comparative genomic analyses are essential in helping to identify protein-coding genes (by homology searching), in evolutionary studies such as analyses of genome duplications, and in helping us to uncover biochemical pathways that allow cells to survive.

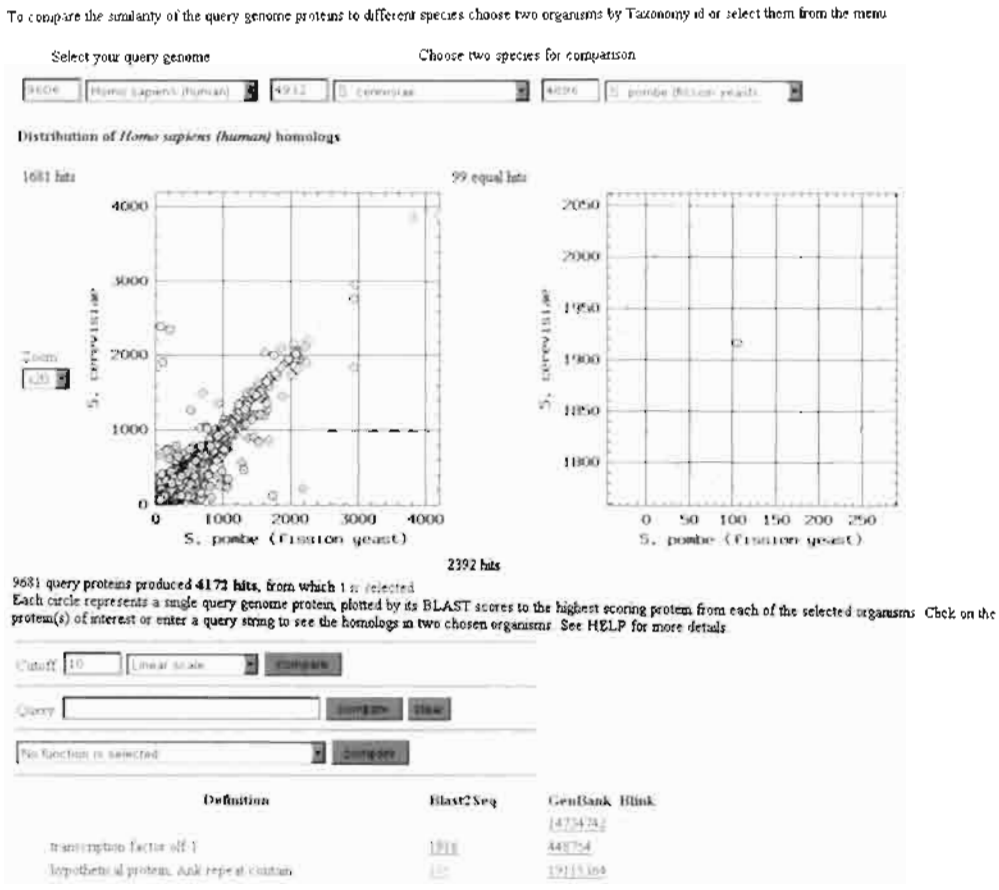


FIGURE 15.23. *The TaxPlot tool at NCBI shows that many* S. cerevisiae *and* S. pombe *proteins are closely related to human proteins but not to each other. The proteomes from many organisms can be compared using this tool. Select a query organism to serve as a reference (in this case,* Homo sapiens*). Then select two organisms to compare; here, we select the fungi* S. cerevisiae *and* S. pombe *from the pull-down menus at top. The main output shows a plot of blastp scores for each human protein against proteins from each fungus. Ninety-nine hits are equal, while 1682 are more closely related to* S. cerevisiae *and 2391 are more closely related to* S. pombe*. It is easy to explore the proteins that are most closely related to a human ortholog in one or the other fungus; for example, by clicking on a data point in the plot at left, we can see that this protein (highlighted in the plot at right and bottom of figure) is transcription factor* olf-1 *(Wang and Reed, 1993) with a score of 1916 in a pairwise comparison of human and* S. cerevisiae *but a score of only 105 in a pairwise comparison of human and* S. pombe.

PITFALLS

At the same time that *S. cerevisiae* serves as an important model organism, it is important to realize the scope of our ignorance. How does the genotype of a single gene knock-out lead to a particular phenotype? We urgently need to answer this question for gene mutations in humans that cause disease, but even in a so-called simple model organism such as yeast we do not understand the full repertoire of protein–protein interactions that underlie cell function. If we think of the genome as a blueprint of a machine, we now have a "parts list" in the form of a list of the gene products. We must next figure out how the parts fit together to allow the machine to function in a variety of contexts. Gene annotation in yeast databases such as SGD and MIPS, including the results of broad functional genomics screens, provides an excellent starting point for functional analyses.

WEB RESOURCES

The SGD (http://genome-www.stanford.edu/Saccharomyces/yeast_info.html) lists a series of yeast resources. Another useful gateway is the Virtual Library–Yeast (http://genome-www.stanford.edu/Saccharomyces/VL-yeast.html).

TABLE 15-10 Fungus Resources

Description	URL
S. pombe genome-sequencing project	http://www.sanger.ac.uk/Projects/S_pombe/
	http://aspergillus-genomics.org/
	http://sequence-www.stanford.edu/group/candida/index.html
C. albicans information	http://alces.med.umn.edu/Candida.html
Fungal genome resource home	http://gene.genetics.uga.edu/
S. cerevisiae promoter database	http://cgsigma.cshl.org/jian/
Tools for regulatory sequence analysis	http://www.ucmb.ulb.ac.be/bioinformatics/rsa-tools/
SAGE advanced query	http://genome-www.stanford.edu/Saccharomyces/SAGE/AdvancedQuery.html
Yeast SAGE	http://www.sagenet.org/yeast/yeastintro.htm
EUROFAN projects	http://www.mips.biochem.mpg.de/proj/eurofan/index.html
Yale genome analysis center	http://ygac.med.yale.edu/
Cell cycle analysis project	http://genome-www.stanford.edu/cellcycle/
	http://cmgm.stanford.edu/pbrown/explore/index.html
Genomewide expression page	http://web.wi.mit.edu/young/expression/
Transcriptional program of sporulation	http://cmgm.stanford.edu/pbrown/sporulation/
Fungal genetics stock center	http://www.fgsc.net/
Yeast virtual library	http://genome-www.stanford.edu/Saccharomyces/VL-yeast.html
	http://www.bio.net:80/hypermail/YEAST/
Munich information center for protein sequences home	http://mips.gsf.de/
MIPS yeast genome	http://www.mips.biochem.mpg.de/proj/yeast/
	http://www.mips.biochem.mpg.de/cgi-bin/proj/expression/start.pl
Saccharomyces genome database	http://genome-www.stanford.edu/Saccharomyces/
	http://www.proteome.com/databases/
Mitochondrial biogenesis database	

DISCUSSION QUESTIONS

[15-1] The budding yeast *Saccharomyces cerevisiae* is sometimes described as a simple organism because it is unicellular, its genome encodes a relatively small number of genes (about 6000), and it has been served as a model organism for genetics studies. Still, we understand the function of only about half its genes. Many functional genomics tools are now available, such as a complete collection of yeast knock-out strains (i.e., null alleles of each gene). How would you use such functional genomics tools to further our knowledge of gene function in yeast?

[15-2] The fungi are a sister group to the metazoans (animals) (Fig. 15.1). Do you expect the principles of genome evolution, gene function, and comparative genomics that are elucidated by studies of fungi to be closely applicable to metazoans such as humans, worms, and flies? For example, we discussed the whole-genome duplication of some

fungi; do you think the human genome also underwent a similar duplication? In comparative genomics, do you expect fungi to be far more similar to each other in their biological properties than metazoans are to each other?

PROBLEMS

[15-1] How many lipocalins do yeast have?

[15-2] ABC transporters constitute a large family of transmembrane-spanning proteins that hydrolyze ATP and drive the transport of ligands such as chloride across a membrane. How many ABC transporters are there in yeast?

[15-3] Use the *Saccharomyces* Genome Database:

- Go to the SGD site (http://genome-www4.stanford.edu/cgi-bin/SGD/seqTools).
- Pick any uncharacterized ORF. To find one, use the Gene/Seq Resources (one of the analysis tools), pick a chromosome (e.g, XII), then select Chromosomal Features Table. The first hypothetical ORF listed is YLL067C.
- Try to explore what its function might be. For some uncharacterized ORFs there will be relatively little information available; for others you may find a lot. From the Chromosomal Features Table click "Info" to view a page similar to that shown in Figure 15.12.
  - -What are the physical properties of the protein (e.g., molecular weight, isoelectric point)?
  - -Does the protein have known domains?
  - -Have interactions been characterized between this and other proteins?
  - -Is the gene either induced or repressed in various physiological states, such as stress response or during sporulation?
- Try using Function Junction (at the bottom of the information page for your ORF). This will simultaneously search six databases:
  - -Yeast Path Calling (two-hybrid analysis)
  - -SGD SAGE query
  - -Worm-Yeast protein comparison
  - -Yeast Microarray Global Viewer
  - -Yeast Protein Function Assignment
  - -Triples database
- In what other organisms is this gene present? Compare the usefulness of exploring SGD versus performing your own BLAST searches to answer this question. Which is better?

SELF-TEST QUIZ

[15-1] You identify a yeast protein in the SGD (*Saccharomyces* Genome Database). It is from ORF YDR477W. What is the *best* place you can go next to find more information about this protein?
(a) MIPS
(b) COG (Clusters of Orthologous Groups)
(c) PDB (Protein Data Bank)
(d) GenBank
(e) KEGG (Kyoto Encyclopedia of genes and Genomes)

[15-2] The yeast *Saccharomyces cerevisiae* is an attractive model organism for many reasons. Which one of the following is NOT a useful feature of yeast?
(a) The genome size is relatively small.
(b) Gene knock-outs by homologous recombination are possible.
(c) Large repetitive DNA sequences serve as a good model for higher eukaryotes.
(d) There is high open reading frame (ORF) density.

[15-3] *Saccharomyces* Genome Database can do all of the following EXCEPT:
(a) Show the potential protein-protein interactions between many yeast proteins
(b) Display chromosomes and all their putative ORFs
(c) Provide a list of bacterial orthologs to yeast proteins
(d) Provide microarray gene expression data for yeast genes

[15-4] The *Saccharomyces cerevisiae* genome is small (it encodes about 6000 genes). It is thought that, about 100 MYA:
(a) The entire genome duplicated, followed by tetraploidization.
(b) The genome underwent many segmental duplications, followed by gene loss.
(c) The entire genome duplicated, followed by gene loss.
(d) The genome duplicated, followed by gene conversion.

[15-5] After gene duplication, the most common fate of the duplicated gene pair is that:
(a) Both copies persist in the genome, maintaining similar functions.
(b) One copy is deleted.
(c) One copy acquires mutations and becomes a pseudogene.
(d) One or both genes diverge functionally.

[15-6] While there are many definitions of "functional genomics," select the best of these choices:

(a) The assignment of function to genes based primarily on genomewide gene expression data using techniques such as microarrays or SAGE
(b) The assignment of function to genes based primarily on comprehensive surveys of protein–protein interactions and protein networks
(c) The combined use of genetic, biochemical, and cell biological approaches to study the function of a gene, its mRNA product, and its corresponding protein product
(d) The assignment of function to genes and proteins using genomewide screens and analyses

[15-7] A major advantage of genetic footprinting using transposons (pioneered by Patrick Brown) is:
(a) The approach is technically easy and can be scaled up to study the function of many genes.
(b) Both insertion alleles and knock-out alleles can be studied.
(c) Any known gene of interest can be studied with this approach.
(d) Mutant strains can be banked for later study by other researchers.

[15-8] The "YKO" project is an effort to systematically knock out all yeast ORFs. A potential limitation of this approach is:
(a) Molecular barcodes may sometimes be toxic for yeast genes.
(b) This approach is not suited to finding new genes but instead focuses on already known genes.
(c) Mutant knock-out strains cannot be banked for later study by other investigators.
(d) Mutations may not be null.

[15-9] One of the most remarkable features of the *Schizosaccharomyces pombe* genome is that:
(a) It is predicted to encode fewer than 5000 proteins, making its genome (and proteome) smaller than even some bacterial genomes.
(b) The number of predicted introns is about the same as the number of predicted ORFs.
(c) It has as many genes that are homologous to bacterial genes as it has genes that are homologous to *S. cerevisiae* genes.
(d) Its genome size is approximately the same as that of *S. cerevisiae,* even though these species diverged hundreds of millions of years ago.

[15-10] Yeast is the only major research organism approved by the U.S. Food and Drug Administration (FDA) for human consumption.
(a) True
(b) False

Suggested Reading

A superb overview of fungal taxonomy is provided by Guarro and Stchigel (1999). This review also introduces fungi from a clinical perspective. A comprehensive taxonomy of the eukaryotes by Baldauf et al. (2000) will be discussed in detail in Chapter 16 (Figure 16.1). Here, we excerpted part of that phylogeny (Fig. 15.1) to introduce the fungi and their close relationship to animals.

An excellent overview of the *S. cerevisiae* genome is provided by Johnston (2000). This article discusses strategies to assign functions to yeast genes.

References

Aalto, M. K., Ronne, H., and Keranen, S. Yeast syntaxins Sso1p and Sso2p belong to a family of related membrane proteins that function in vesicular transport. *EMBO J.* **12,** 4095–4104 (1993).

Ainsworth, G. C. Fungus infections (mycoses). In K. F. Kiple, ed. *The Cambridge World History of Human Disease.* Cambridge University Press, New York, 1993, pp. 730–736.

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., and Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290,** 972–977 (2000).

Braun, E. L., Halpern, A. L., Nelson, M. A., and Natvig, D. O. Large-scale comparison of fungal sequence information: Mechanisms of innovation in *Neurospora crassa* and gene loss in *Saccharomyces cerevisiae. Genome Res.* **10,** 416–430 (2000).

Bulloch, W. *The History of Bacteriology.* Oxford University Press, New York, 1938.

Casselton, L., and Zolan, M. The art and design of genetic screens: Filamentous fungi. *Nat. Rev. Genet.* **3,** 683–697 (2002).

Cherry, J. M., et al. Genetic and physical maps of *Saccharomyces cerevisiae. Nature* **387,** 67–73 (1997).

Chu, S., et al. The transcriptional program of sporulation in budding yeast. *Science* **282,** 699–705 (1998).

Dacks, J. B., and Doolittle, W. F. Novel syntaxin gene sequences from *Giardia, Trypanosoma* and algae: implications for the ancient evolution of the eukaryotic endomembrane system. *J. Cell Sci.* **115,** 1635–1642 (2002).

DeRisi, J. L., Iyer, V. R., and Brown, P. O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278,** 680–686 (1997).

DeRisi, J., et al. Genome microarray analysis of transcriptional activation in multidrug resistance yeast mutants. *FEBS Lett* **470,** 156–160 (2000).

Dujon, B., et al. Complete DNA sequence of yeast chromosome XI. *Nature* **369,** 371–378 (1994).

Dwight, S. S., et al. Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.* **30,** 69–72 (2002).

Garbarino, J. E., and Gibbons, I. R. Expression and genomic analysis of midasin, a novel and highly conserved AAA protein distantly related to dynein. *BMC Genomics* **3,** 18 (2002).

Giaever, G., et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418,** 387–391 (2002).

Goffeau, A., et al. Life with 6000 genes. *Science* **274,** 546, 563–567 (1996).

Guarro, J., Gene J., and Stchigel, A. M. Developments in fungal taxonomy. *Clin. Microbiol. Rev.* **12,** 454–500 (1999).

Harrison, P. M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. A question of size: The eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.* **30,** 1083–1090 (2002).

Holstege, F. C., et al. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* **95,** 717–728 (1998).

Jelinsky, S. A., and Samson, L. D. Global response of *Saccharomyces cerevisiae* to an alkylating agent. *Proc. Natl. Acad. Sci. USA* **96,** 1486–1491 (1999).

Johnston, M. The yeast genome: On the road to the Golden Age. *Curr. Opin. Genet. Dev.* **10,** 617–623 (2000).

Johnston, M., et al. The nucleotide sequence of *Saccharomyces cerevisiae* chromosome XII. *Nature* **387,** 87–90 (1997).

Katinka, M. D., et al. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi. Nature* **414,** 450–453 (2001).

Kinsey, J. A., Garrett-Engele, P. W., Cambareri, E. B., and Selker, E. U. The *Neurospora* transposon Tad is sensitive to repeat-induced point mutation (RIP). *Genetics* **138,** 657–664 (1994).

Krumlauf, R., and Marzluf, G. A. Genome organization and characterization of the repetitive and inverted repeat DNA sequences in *Neurospora crassa. J. Biol. Chem.* **255,** 1138–1145 (1980).

Kuchenmeister, F. *On animal and vegetable parasites of the human body, a manual of their natural history, diagnosis, and treatment.* Sydenham Society, London, 1857.

Kumar, A., et al. An integrated approach for finding overlooked genes in yeast. *Nat. Biotechnol.* **20,** 58–63 (2002a).

Kumar, A., et al. The TRIPLES database: A community resource for yeast molecular biology. *Nucleic Acids Res.* **30,** 73–75 (2002b).

Lowe, T. M., and Eddy, S. R. A computational screen for methylation guide snoRNAs in yeast. *Science* **283,** 1168–1171 (1999).

Lynch, M., and Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290,** 1151–1155 (2000).

Mackiewicz, P., et al. How many protein-coding genes are there in the *Saccharomyces cerevisiae* genome? *Yeast* **19,** 619–629 (2002).

Marc, P., Devaux, F., and Jacq, C. yMGV: A database for visualization and data mining of published genome-wide yeast expression data. *Nucleic Acids Res.* **29,** E63–3 (2001).

Margulis, L., and Schwartz, K. V. *Five Kingdoms. An Illustrated Guide to the Phyla of Life on Earth.* W. H. Freeman and Company, New York, 1998.

Mewes, H. W., et al. Overview of the yeast genome. *Nature* **387,** 7–65 (1997).

Mewes, H. W., et al. MIPS: A database for genomes and protein sequences. *Nucleic Acids Res.* **30,** 31–34 (2002).

Ohno, S. *Evolution by Gene Duplication.* Springer Verlag, Berlin, 1970.

Oshiro, G., et al. Parallel identification of new genes in *Saccharomyces cerevisiae. Genome Res.* **12,** 1210–1220 (2002).

Perkins, D. D., Radford, A., and Sachs, M. S. *The Neurospora Compendium: Chromosomal loci.* Academic Press, San Diego, CA, 2001.

Piskur, J. Origin of the duplicated regions in the yeast genomes. *Trends Genet.* **17,** 302–303 (2001).

Protopopov, V., Govindan, B., Novick, P., and Gerst, J. E. Homologs of the synaptobrevin/VAMP family of synaptic vesicle proteins function on the late secretory pathway in *S. cerevisiae. Cell* **74,** 855–861 (1993).

Ross-Macdonald, P., et al. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402,** 413–418 (1999).

Roth, J. F. The yeast Ty virus-like particles. *Yeast* **16,** 785–795 (2000).

Rozen, S., et al. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* **423,** 873–876 (2003).

Sankoff, D. Gene and genome duplication. *Curr. Opin. Genet. Dev.* **11,** 681–684 (2001).

Selker, E. U. Premeiotic instability of repeated sequences in *Neurospora crassa. Annu. Rev. Genet.* **24,** 579–613 (1990).

Seoighe, C., and Wolfe, K. H. Updated map of duplicated regions in the yeast genome. *Gene* **238,** 253–261 (1999).

Smith, V., Botstein, D., and Brown, P. O. Genetic footprinting: A genomic strategy for determining a gene's function given its sequence. *Proc. Natl. Acad. Sci. USA* **92,** 6479–6483 (1995).

Smith, V., Chou, K. N., Lashkari, D., Botstein, D., and Brown, P. O. Functional analysis of the genes of yeast chromosome V by genetic footprinting. *Science* **274,** 2069–2074 (1996).

Spellman, P. T., et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.* **9,** 3273–3297 (1998).

Wagner, A. Robustness against mutations in genetic networks of yeast. *Nat. Genet.* **24,** 355–361 (2000).

Wagner, A. Birth and death of duplicated genes in completely sequenced eukaryotes. *Trends Genet.* **17,** 237–239 (2001).

Wang, D. Y., Kumar, S., and Hedges, S. B. Divergence time estimates for the early history of animal phyla and the origin of

plants, animals and fungi. *Proc. R. Soc. Lond. B Biol. Sci.* **266,** 163–171 (1999).

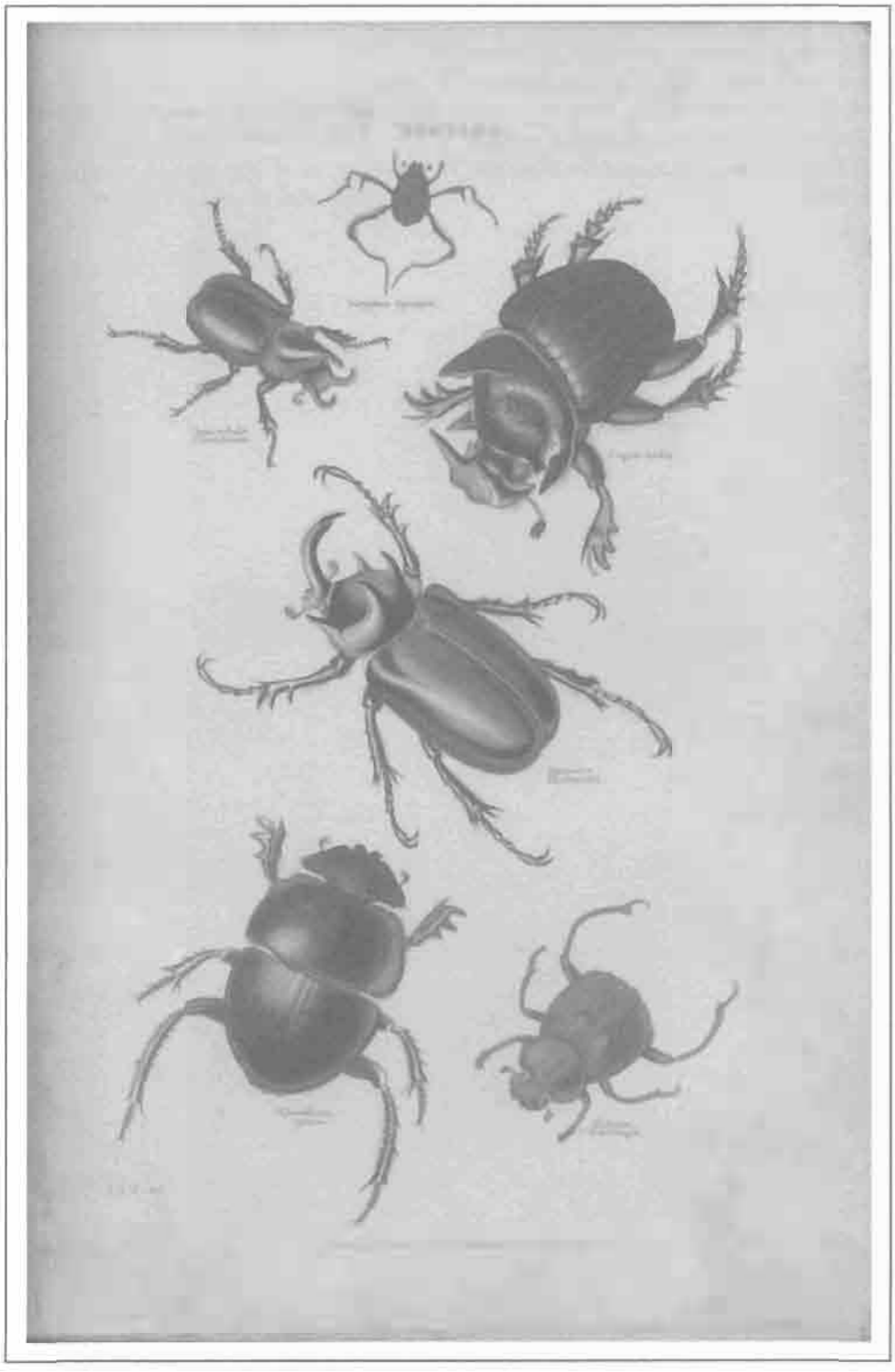
Wang, M. M., and Reed, R. R. Molecular cloning of the olfactory neuronal transcription factor Olf-1 by genetic selection in yeast. *Nature* **364,** 121–126 (1993).

Winzeler, E. A., et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285,** 901–906 (1999).

Wolfe, K. H., and Shields, D. C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387,** 708–713 (1997).

Wong, S., Butler, G., and Wolfe, K. H. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. *Proc. Natl. Acad. Sci. USA* **99,** 9272–9277 (2002).

Wood, V., et al. The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415,** 871–880 (2002).



In Chapter 16, we discuss the dazzling variety of eukaryotes from a bioinformatics perspective. Baron Georges Cuvier (1769–1832) attempted a systematic classification of animals. He described four great divisions of the animal kingdom: vertebrate animals, molluscous animals, articulate animals, and radiate animals (also called Zoophytes). This figure depicts several species of beetles (Cuvier, 1849, plate XXXI). Since the 18th century, 350,000 species of beetles have been described, but by some estimates there are many millions of beetle species (Wilson, 1992).

16

Eukaryotic Genomes: From Parasites to Primates

INTRODUCTION

The eukaryotes are single-celled or multicellular organisms that are characterized by the presence of a membrane-bound nucleus and a cytoskeleton. We began our examination of eukaryotes with the fungi (Chapter 15), including *Saccharomyces cerevisiae*. In this chapter we will broadly survey the eukaryotes, from the simplest primitive single-celled organisms to plants and metazoans (animals).

In the first half of this chapter, we will explore general features of eukaryotic genomes. We will examine genome sizes, noncoding DNA (e.g., repetitive DNA), and coding DNA (genes). For a given segment of genomic DNA, we will address the problem of annotation: how much noncoding DNA is present and of what type? How many protein-coding genes or RNA genes are present? What is the definition of a gene, and what software programs are available to detect genes?

More broadly, as we examine genomic DNA, we can address the issue of how organisms and species evolve. How are the genomes of different eukaryotes organized into chromosomes? What are the differences between the mouse and rat genomes? How do new species of mosquito emerge from those that are alive today? In addition

Synonyms of eukaryotes include eucaryotae, eucarya, eukarya, and eukaryotae. The word derives from the Greek *eu-* ("true") and *karutos* ("having nuts"; this refers to the nucleus).

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

FIGURE 16.1. *A phylogeny of eukaryotes based on parsimony analysis of concatenated protein sequences. The proteins analyzed were EF-1α (abbreviated E in tree), actin (C), α-tubulin (A), and β-tubulin (B). This tree may be compared to the eukaryotic portion of the global tree of life based upon small-subunit ribosomal RNA sequences (Fig. 12.1). In this tree, 14 kingdoms are indicated as well as seven supergroups. One of the supergroups, Opisthokonta, includes fungi and microsporidia (Chapter 15) and metazoa (vertebrate and invertebrate animals). The tree was constructed by maximum parsimony (with bootstrap values indicated above the horizontal branches) and by maximum-likelihood analysis of second-codon-position nucleotides. For taxa with missing data, the sequences used are indicated in brackets (e.g., [EAB]). Modified from Baldauf et al. (2000); see their Figures 1 and 2 for further details. Used with permission.*

to our focus on coding DNA, noncoding DNA, and proteins, we will address eukaryotic chromosome structure and function.

In the second half of this chapter we will explore individual eukaryotic genomes, from parasites to primates. We will refer to a phylogenetic tree of the eukaryotes that was produced by Baldauf et al. (2000) (Fig. 16.1). This tree was created by parsimony analysis using four concatenated protein sequences: elongation factor 1a (EF-1α), actin, α-tubulin, and β-tubulin. We already discussed fungal genomes in Chapter 15; they are represented in a group that is adjacent to the metazoa (animals). We will examine representative organisms in this tree, moving from the bottom up. This includes the diplomonad *Giardia lamblia* and other protozoans, such as the

malaria parasite, *Plasmodium falciparum*; the plants, including the first sequenced plant genome (that of the thale cress, *Arabidopsis thaliana*) and rice (*Oryza sativa*); and the metazoans, from worms and insects to fish and mammals. We will address the human genome in Chapter 17.

A focus of this chapter is on the complete sequencing of eukaryotic genomes. Why should we want to obtain the complete genomic sequences (The C. elegans sequencing consortium, 1998)?

- The complete genome sequence provides the basis for discovering all the genes that are encoded in a genome. Other approaches, such as characterizing expressed sequence tags, can never be as comprehensive.
- The comparative genomic sequence shows the structural and regulatory elements associated with genes.
- It provides the basis to assess the molecular evolution of a species as well as the extent of its variation between individuals, populations, and other species.
- It provides a set of tools for future experimentation.

At the same time that eukaryotic genomes are completely sequenced, a parallel molecular approach is the characterization of individual genes from many hundreds or thousands of species. The use of model genes complements the use of model organisms. For example, a search of GenBank for ribulose-1,5-bisphosphate carboxylase (rubisco; *rbcL*) currently reveals over 15,000 entries. Rubisco is a major plant gene (discussed below), and the availability of molecular sequence data from many species for this and selected other genes is crucial for phylogenetic reconstructions and structure–activity studies. Other commonly studied genes include highly conserved molecules such as actin, tubulin, and those described in Table 11.2.

General Features of Eukaryotes

Major Differences between Eukaryotes and Prokaryotes

Eukaryotes share a common ancestry with prokaryotes, but when we compare them, we find several outstanding differences (Cavalier-Smith, 2002; Vellai and Vida, 1999; Watt and Dean, 2000). Some of these genomic features are highlighted in Table 16.1.

- There is a tremendous diversity of both prokaryotic and eukaryotic life forms. However, very few bacterial or archaeal life forms are visible to the human eye. Many eukaryotes are single-celled, microscopic organisms. Nonetheless, we tend to associate eukaryotes with multicellular organisms (e.g., plants and metazoans), in contrast to the prokaryotes.
- Eukaryotic cells have three cellular features that are lacking in prokaryotes: (1) a membrane-bound nucleus, (2) an extensive system of organelles bound by intracellular membranes, and (3) a cytoskeleton, including elements such as actin and tubulin, and molecular motors. Notably, prokaryotes lack energy-producing organelles and are incapable of endocytosis, the process by which extracellular cargo is internalized (Vellai and Vida, 1999).
- Most eukaryotes undergo sexual reproduction, although some are asexual. Bacteria lack gamete fusion and do not exchange DNA by sex.

Sexual reproduction is called syngamy, the process by which the haploid chromosomes of the male and female gametes combine to form the zygote (i.e., the fertilized ovum).

TABLE 16-1 Features of Several Sequenced Bacterial and Eukaryotic Genomes

Feature	*E. coli* K-12	Parasite[a]	Yeast[b]	Slime Mold[c]	Plant[d]	Human[e]
Genome size, Mb	4.64	22.8	12.5	8.1	115	3289
GC content, %	50.8	19.4	38.3	22.2	34.9	41
Number of genes	4288	5268	5770	2799	25,498	30,000–40,000
Gene density, kb per gene	0.95	4.34	2.09	2.60	4.53	27
Percent coding	87.8	52.6	70.5	56.3	28.8	1.3
Number of introns	0	7406	272	3578	107,784	53,295

Source: Adapted from Gardner et al. (2002); Blattner et al. (1997); International Human Genome Sequencing Consortium (2001).
[a] *Plasmodium falciparum.*
[b] *Saccharomyces cerevisiae.*
[c] *Dictyostelium discoideum.*
[d] *Arabidopsis thaliana.*
[e] *Homo sapiens.*
Abbreviations: bp, base pairs; Mb, millions of base pairs (megabases).

TABLE 16-2 Genome Size of Selected Eukaryotes

Taxon	Phylum, Class, or Division	Genome Size Range (Gb)	Ratio of genome sizes (Highest/Lowest)
All eukaryotes	–	0.003–686	228,667
Alveolata	–	–	22,333
	Apicomplexians	0.009–201	22,333
	Ciliates	0.024–8.62	359
	Dinoflagellates	1.37–98	72
Diatoms		0.035–24.5	700
Amoebae		0.035–686	19,600
Euglenozoa		0.098–2.35	24
Fungi/microsporidia		0.003–1.47	490
Animals	–	–	3,325
	Sponges	0.059–1.78	30
	Cnidarians	0.227–1.83	8
	Insects	0.089–9.47	106
	Elasmobranchs	1.47–15.8	11
	Bony fishes	0.345–133	386
	Amphibians	0.93–84.3	91
	Reptiles	1.23–5.34	4
	Birds	1.67–2.25	1
	Mammals	1.7–6.7	4
	Placozoa	0.04	–
Plants	–	–	6,140
	Algae	0.080–30	375
	Pteridophytes	0.098–307	3,133
	Gymnosperms	4.12–76.9	19
	Angiosperms	0.050–125	2,500

Source: Adapted from Graur and Li (2000); Animal Genome Size Database of T. R. Gregory (▶http://www.genomesize.com) and the National Center for Biotechnology Information (▶http://www.ncbi.nlm.nih.gov).
Note: 0.001 Gb (gigabases) equals 1 Mb. Values in picograms were multiplied times 0.9869×10^9 to obtain gigabases.

- The genome size of eukaryotes varies widely, spanning five orders of magnitude (Table 16.2). In contrast, most archaeal and bacterial genomes are between 0.5 and about 13 Mb in size (see Chapters 12 and 14).
- Prokaryotic genomes tend to have a relatively high density of protein-coding genes and little repetitive or other noncoding DNA. For example, 0.7% of the *Escherichia coli* genome consists of noncoding repeats (Blattner et al., 1997). In contrast, the majority of many eukaryotic genomes are composed of noncoding DNA. Several examples are provided in Table 16.1.
- Prokaryotes are haploid. Eukaryotes may be haploid or diploid or have other ploidy states. This higher level of ploidy offers eukaryotes a variety of evolutionary mechanisms such as heterozygous advantage (Watt and Dean, 2000).
- The genomes are organized differently. Eukaryotic nuclear genomes are organized primarily into linear chromosomes. (The majority of bacterial and archaeal genomes are circular; see Fig. 14.1.) These eukaryotic chromosomes have a centromere as well as telomeres at either end. These features are absent from prokaryotic chromosomes, although centromere-like elements have been described (Ben-Yehuda et al., 2002; Moller-Jensen et al., 2002). The mechanisms by which bacteria segregate DNA are relatively obscure.

C Value Paradox: Why Eukaryotic Genome Sizes Vary So Greatly

In eukaryotic genomes, the haploid genome size (C value) varies enormously. This is shown in Table 16.2 for various taxa of eukaryotes and in Table 16.3 for specific eukaryotic species. Some genomes are relatively quite small, such as the microsporidian *Encephalitozoon cuniculi* (2.9 Mb; Chapter 15). Others have genome sizes in the range of hundreds of billions of base pairs. Tremendous variation in C values occurs among the unicellular protists such as amoebae, with a 20,000-fold range. Within the animal kingdom, the range is about 3000-fold.

Remarkably, the range in C values does not correlate well with the complexity of organisms. Some organisms such as *A. thaliana* (a plant) and *Fugu rubripes* (a pufferfish) have extremely compact genomes while closely related organisms of similar biological complexity have genomes that are orders of magnitude larger. This lack of correlation is called the C value paradox (Hartl, 2000; Knight, 2002; Hancock, 2002; Kidwell, 2002). The genomes of several eukaryotes were sequenced in either a draft or finished version in recent years, including *Caenorhabditis elegans* (1998), *Drosophila melanogaster* (2000), *Homo sapiens* (2001), and *Mus musculus* (2002) (see below and Chapter 12). These whole-genome studies provide one clear answer to the C value paradox: Genomes are filled with large tracts of noncoding DNA sequences in varying amounts. This accounts for the variation in genome size. We will explore this noncoding DNA next.

Many Eukaryotic Genomes Consist of Noncoding and Repetitive DNA Sequences

Bacterial and archaeal genomes have both genes and additional, relatively small intergenic regions. Typically, these prokaryotic genomes are circular, and there is about one gene in each kilobase of genomic DNA (Chapter 14 and Table 16.1). In contrast, eukaryotic genomes contain a smaller proportion of protein-coding genes and large amounts of noncoding DNA. This noncoding material includes repetitive DNA,

The C value is measured in base pairs or in picograms (pg) of DNA. One picogram of DNA corresponds to approximately 1 Gb.

In eukaryotes, normal germ cells are haploid while somatic cells are usually diploid. Thus, different cells within an individual can have different ploidy. Ploidy is the number of chromosome sets in a cell.

Ploidy can vary in many ways. Some single-celled eukaryotes such as *S. cerevisiae* can grow in either the haploid or diploid state. Triploid *Drosophila* are viable (but with reduced fertility). Although we distinguish the ploidy state in germ cells and somatic cells, ploidy can also vary in somatic cells within an individual. For example, in humans a small fraction of liver cells is typically triploid. In general an extra copy of even one chromosome is usually lethal in mammals (see below).

An online database of plant C values (by M. D. Bennett and I. J. Leitch) is available at ▸http://www.rbgkew.org.uk/cval/homepage.html. The Animal Genome Size Database (from T. Ryan Gregory) is online at ▸http://www.genomesize.com/. Another resource is the Database of Genome Size (DOGS) (▸http://www.cbs.dtu.dk/databases/DOGS/).

TABLE 16-3 Genome Size (*C* Value) for Various Eukaryotes

Species	Common Name	C value (Gb)
Saccharomyces cerevisiae	Yeast	0.012
Neurospora crassa	Fungus	0.043
Dysidea crawshagi	Sponge	0.054
Caenorhabditis elegans	Nematode	0.097
Drosophila melanogaster	Fruitfly	0.12
Paramecium aurelia	Ciliate	0.19
Oryza sativa	Rice	0.47
Strongylocentrotus purpuratus	Sea urchin	0.80
Gallus domesticus	Chicken	1.23
Erysiphe cichoracearum	Powdery mildew	1.5
Boa constrictor	Snake	2.1
Parascaris equorum	Roundworm	2.5
Carcharias obscurus	Sand-tiger shark	2.7
Canis familiaris	Dog	2.9
Rattus norvegicus	Rat	2.9
Xenopus laevis	African clawed frog	3.1
Homo sapiens	**Human**	**3.3**
Nicotania tabacum	Tobacco plant	3.8
Locusta migratoria	Migratory locust	6.6
Paramecium caudatum	Ciliate	8.6
Allium cepa	Onion	15
Truturus cristatus	Warty newt	19
Thuja occidentalis	Western giant cedar	19
Coscinodisucus asteromphalus	Centric diatom	25
Lilium formosanum	Lily	36
Amphiuma means	Two-toed salamander	84
Pinus resinosa	Canadian red pine	68
Protopterus aethiopicus	Marbled lungfish	140
Amoeba proteus	Amoeba	290
Amoeba dubia	Amoeba	690

Source: Adapted from Graur and Li (2000); NCBI (▶http://www.ncbi.nlm.nih.gov); Cameron et al. (2000); and the Database of Genome Sizes (▶http://www.cbs.dtu.dk/databases/DOGS/abbr.table.common.txt).

Alternative solutions to the C value paradox do not fit. The number of protein-coding genes in eukaryotes varies over a ≈10-fold range, but this variation is far smaller than the range of genome sizes. Also, interspecies variation in the lengths of mRNA molecules does not explain the *C* value paradox because no correlation exists between mean gene length and genome size.

genes encoding RNAs that have regulatory functions, and introns that interrupt exons and are spliced from mature RNA transcripts.

A large portion of the noncoding sequences in eukaryotes consists of repetitive DNA sequences. These are repeated nucleotides of various lengths that occur throughout the genome (Jurka, 1998). We will also discuss these repeats in our analysis of the human genome (Chapter 17). In mammals, up to 60% of genomic DNA is repetitive; in some yeasts 20% is repetitive.

Britten and Kohne (1968) performed some of the earliest experiments that defined the repetitive nature of eukaryotic DNA. They purified genomic DNA from a wide variety of species, sheared it, and dissociated the DNA strands. Under

FIGURE 16.2. *The complexity of genomic DNA can be estimated by denaturing then renaturing DNA. This figure (redrawn from Britten and Kohne, 1968) depicts the relative quantity of mouse genomic DNA (y axis) versus the logarithm of the frequency with which the DNA is repeated. The data are derived from a C_0t curve, which describes the percent of genomic DNA that reassociates at particular times and DNA concentrations. A large $C_0t_{1/2}$ value implies a slower reassociation reaction. Three classes are apparent. The fast component accounts for 10% of mouse genomic DNA (arrow A), and it represents highly repetitive satellite DNA. An intermediate component accounts for about 20% of mouse genomic DNA and contains repeats having from 1000 to 100,000 copies. The slowly reassociating component, comprising 70% of the mouse genome, corresponds to unique, single-copy DNA. Britten and Kohne (1968) obtained similar profiles from other eukaryotes, although distinct differences were evident between species. Used with permission.*

appropriate conditions of salt, temperature, and time, the DNA strands reanneal. They measured the rate at which the DNA reasssociates and found that for dozens of eukaryotes—but not for several viruses or bacteria—DNA reassociates in several distinct fractions. Large amounts of eukaryotic DNA reassociate extremely rapidly. For the mouse genome, about 10% of genomic DNA reassociates rapidly and consists of about one million copies (Fig. 16.2, arrow A). This highly repetitive DNA is localized to the highly condensed portion of chromosomes referred to as heterochromatin (Redi et al., 2001; Avramova, 2002). A further 20% of the DNA reassociates in a fraction containing from 1000 to 100,000 distinct DNA species (arrow B). Finally, about 70% of the DNA is unique, consisting of only a single copy (arrow C). This DNA forms the euchromatin, a portion of the chromosome that is not uncondensed and thus is accessible for the transcription of genes. The banding pattern of chromosomes (Fig. 16.19 below) corresponds to regions of heterochromatin and euchromatin.

The origin of these repeats and their function present fascinating questions. What different kinds of repeats occur? From where did they originate and when? Is there a logic to their promiscuous growth or do they multiply without purpose? We are beginning to understand the extent and nature of the repeat content of eukaryotic genomes, including the human genome. Repetitive DNA has sometimes been called "junk DNA" or "selfish DNA," reflecting its propensity to expand throughout genomes. However, it is likely that repetitive DNA has important roles in chromosome structure, recombination events, and the function of some genes [Makalowski (2000) and see below].

Britten and Kohne (1968) used several techniques to distinguish single-stranded from double-stranded DNA, such as hydroxyapatite chromatography (a calcium phosphate column), binding of radiolabeled DNA fragments to immobilized DNA on filters, and spectrophotometry. The rate of DNA reassociation is a function of the incubation time (t) and the DNA concentration (C_0). The C_0t plot displays the fraction of DNA that remains single stranded versus the C_0t value, and it is the basis for the data shown in Figure 16.2.

A retrotransposon (also called a retroposon or retroelement) is a transposable element that copies itself to genomic locations through a process of reverse transcription with an RNA intermediate. This process is similar to that of a retrovirus.

TABLE 16-4 Examples of Classes and Transposable Elements

Class	Subclass	Superfamily	Examples of Family	Approximate Size Range (bp)
Retroelements (RNA-mediated elements)	LTR retrotransposons	Ty1-copia	Opie-1 (maize)	3000–12,000
	Non-LTR retrotransposons	LINEs	*LINE-1* (human)	1000–7000
		SINEs	*Alu* (human)	100–500
DNA transposons	Cut-and-paste transposition	Mariner-Tc1	*Tc1* in *C. elegans*	1000–2000
		P	*P* in *Drosophila*	500–4600
	Rolling circle transposition	*Helitrons*	*Helitrons* in *A. thaliana*, *O. sativa*, and *C. elegans*	5500–17,500

Source: Adapted from Kidwell (2002). Used with permission.

Barbara McClintock was awarded a Nobel Prize in 1983 for her discovery of mobile genetic elements in maize (*Zea mays*). You can read more about this pioneering work at http://www.nobel.se/medicine/laureates/1983/.

Wojciech Makalowski offers a website that describes many examples of SINEs and LINEs as well as hypotheses about their possible functions (http://www.ncbi.nlm.nih.gov/Makalow/sines.html).

A search of Entrez nucleotide with the term "retropseudogene" yields 65 hits (January 2003), while "retrotransposed" yields 10 hits. But a search with the term "retrotransposon" yields over 11,000 matches.

There are five main classes of repetitive DNA in eukaryotes (IHGSC, 2001; Jurka, 1998; Kidwell, 2002; Makalowski, 2000):

1. **Interspersed Repeats (Transposon-Derived Repeats).** Together, interspersed repeats constitute about 45% of the human genome (see Chapter 17). These repeats can be generated by elements that copy RNA intermediates (retroelements) or DNA intermediates (DNA transposons) (Table 16.4). Genes may be copied by retrotransposition when an mRNA is reverse transcribed and then integrated into the genome. Such genes can be identified because they usually lack introns, while they do have short direct flanking repeats. Examples of some mammalian retrotransposed genes are presented in Table 16.5.

TABLE 16-5 Examples of Mammalian Genes Generated by Retrotransposition

Retrotransposed genes lack introns, and they often have flanking direct repeats and a polyadenine tail

Retrotransposed Gene			Original Gene				
Name	RefSeq	Chr	Name	RefSeq	Chr	Distribution	Age (MYA)
ADAM20	NM_003814	14q	*ADAM9*	NM_003816	8p	Human, not macaque	<20
Cetn1	NM_004066	18p	*Cetn2*	NM_004344	Xq28	Mammals	>75
Glud2	NM_012084	Xq	*Glud1*	NM_005271	10q	Human, not mouse	<70
Pdha2	NM_005390	4q	*Pdha1*	NM_000284	Xp	Placentals	~70
SRP46	NM_032102	11q	*PR264/SC35*	NM_003016	17q	Human, simians	~89
Supt4h2	NM_011509	10	*Supt4h*	NM_009296	11	Mouse	<70

Source: Adapted from Betrán and Long (2002) (see that article for additional genes and literature references) and from a search of LocusLink and Entrez (NCBI) with the term *retropseudogene*.
Abbreviations: Chr, chromosome; MYA, millions of years ago; *ADAM*, a disintegrin and metalloproteinase; *Cetn*, centrin, EF-hand protein; *Glud*, glutamate dehydrogenase; *Pdha2*, pyruvate dehydrogenase (lipoamide) alpha 2; *Supt4h*, suppressor of Ty 4 homolog (*S. cerevisiae*).

Interspersed repeats can be divided into four categories (Ostertag and Kazazian, 2001; Kidwell, 2002) (see also Figure 17.18):

- Long-terminal-repeat (LTR) transposons, which are RNA-mediated elements. These are also called retrovirus-like elements. LTR transposons have LTRs of several hundred base pairs at either end of the element.
- Long interspersed elements (LINEs), which encode an enzyme with reverse transcriptase activity (and possibly additional proteins). In mammals, LINE1 and LINE2 families are most prevalent.
- Short interspersed elements (SINEs), which are also RNA-mediated elements. *Alu* repeats, found in primates, are well-known examples of SINEs. We will see an example of an *Alu* repeat sequence in Figure 16.8 (below).
- DNA transposons comprise about 3% of the human genome.

2. **Processed Pseudogenes.** These are genes that are not actively transcribed or translated (Harrison and Gerstein, 2002; Echols et al., 2002). They represent genes that were once functional, but they are defined by their lack of protein product. They can be recognized because of the presence of a stop codon or frameshift that interrupts an open reading frame. Pseudogenes can arise through retrotransposition events (i.e., random insertion events mediated by LINEs having reverse transcriptase activity) or following gene duplication and subsequent gene loss.

3. **Simple Sequence Repeats.** These microsatellites (typically from 1 to 6 bp in length) and minisatellites (typically from a dozen to 500-bp repeats) include short sequences such as $(A)_n$, $(CA)_n$, or $(CGG)_n$. (We will see an example of a CA repeat in human genomic DNA in Fig. 16.8, below.) Replication slippage is a mechanism by which simple sequence repeats may occur.

Simple sequence repeats of particular length and composition occur preferentially in different species. For example, $(AT)_n$ is especially common in *A. thaliana*, and $(CT/GA)_n$ occurs preferentially in *C. elegans* (Schlötterer and Harr, 2000). In *Drosophila virilis*, the density and length of microsatellites are considerably greater than in *D. melanogaster* or *H. sapiens* (Schlötterer and Harr, 2000). In humans, simple sequence repeats are of particular interest because they are highly polymorphic between individuals and thus serve as useful genetic markers. Also, the expansion of triplet repeats such as CAG is associated with 14 diseases, including Huntingon's disease (Cummings and Zoghbi, 2000). We will discuss these issues in Chapter 18 (on human disease).

4. **Segmental Duplications.** Consisting of blocks of about 1 to 200 or 300 kilobases (kb) that are copied from one genomic region to another (Bailey et al., 2001), these duplications occur both within and between chromosomes (intra- and interchromosomally). Bailey et al. estimate that the human genome consists of about 3.6% duplicated regions (having a length that is at least 1 kb and sharing 90–98% sequence similarity).

As an example of a segmental duplication, we will consider a cluster of lipocalin genes on human chromosome 9. The lipocalins of all species have been divided into 14 monophyletic clades (Gutierrez et al., 2000). In humans, the lipocalins include at least 10 genes localized to chromosome 9q32-34. Figure 16.3 presents a schematic view of the genomic DNA, including the tear lipocalin (*LCN1*) and odorant-binding protein genes (adapted from Lacazette et al., 2000). Based on their analysis of this genomic region, Lacazette et al. (2000) proposed a model to account for the lipocalin genes and pseudogenes observed today in a portion of chromosome 9q34 (Fig. 16.3,

Mark Gerstein's laboratory offers a website in pseudogenes (▸http://www.pseudogene.org/). This includes a browser and descriptions of pseudogenes in human, worm, fly, yeast, and plant.

The mouse genome contains one functional gene encoding glyceraldehyde 3-phosphate dehydrogenase (*Gapdh*; NM_008084) and at least 400 pseudogenes distributed across 19 chromosomes (Mouse Genome Sequencing Consortium, 2002). Currently (January 2003), the functional *Gapdh* gene is listed as assigned to mouse chromosomes 7 (Mouse Genome Sequencing Consortium, 2002), 6 (LocusLink and UniGene at NCBI), and 5 (Ensembl mouse genome Contig Viewer). The presence of many pseudogenes contributes to the difficulty of assigning correct chromosomal loci.

Some authors define microsatellites as having a length of 1–6 bp, while others suggest 1–12 bp.

To see specific examples of simple sequence repeats, go to Entrez Nucleotide and enter "microsatellite." There are over 50,000 entries from which to choose.

The Tandem Repeats Finder is an online tool that allows you to search a sequence for tandem repeats of up to 2000 bp (▸http://tandem.biomath.mssm.edu/trf/trf.html) (Benson, 1999).

The Case Western Reserve University (CWRU) Duplication Browser allows you to identify segmental duplications in the human genome. It is available at ▸http://humanparalogy.gene.cwru.edu/SDD/, and it uses the University of California Santa Cruz genome browser that we will explore in Chapter 17. This CWRU site offers a database of over 8500 segmental duplications in the human genome.

FIGURE 16.3. *Genes evolve by successive tandem duplications. Lacazette et al. (2000) proposed this model to explain how a hypothetical lipocalin gene (top of figure) could have evolved to account for the extant pattern of genes and pseudogenes determined by sequence analysis of this locus. First, an ancestral lipocalin having seven exons duplicated (step 1) and functionally diverged (step 2). This region, containing two genes, duplicated twice (step 3) after which one gene was deleted (the hypothetical* OBPIIc *gene) and portions of an* LCN1 *gene were deleted (step 4). Finally, several new exons were recruited (step 5). Used with permission.*

bottom). A hypothetical ancestral lipocalin gene had seven exons and six introns (Fig. 16.3, top), a gene structure typical of mammalian lipocalins (Salier, 2000). This gene duplicated by tandem duplication (Fig. 16.3, step 1), after which the two ancestral genes differentiated to assume distinct functions (step 2). This locus then duplicated twice, generating *LCN1* and *OBPII* paralogs (step 3). However, only two *OBPII* genes are present in this locus today (Fig. 16.3, bottom), and the *LCN1* gene is accompanied by two pseudogenes (*LCN1b* and *LCN1c*). Thus partial duplications may

TABLE 16-6 Telomeric Repeat Sequences from Several Eukaryotic Organisms

Organism	Telomeric Repeat	Reference
Arabidopsis thaliana, other plants	TTTAGGG	McKnight et al., 1997
Ascaris suum (nematode)	TTAGGC	Jentsch et al., 2002
Euplotes aediculatus, Euplotes crassus, Oxytricha nova (ciliates)	TTTTGGGG	Jarstfer and Cech, 2002; Shippen-Lentz and Blackburn, 1989; Melek et al., 1994
Giardia duodenalis, Giardia lamblia	TAGGG	Upcroft et al., 1997; Hou et al., 1995
Guillardia theta (cryptomonad nucleomorph)	$[AG]_7AAG_6A$	Douglas et al., 2001
Homo sapiens, other vertebrates	TTAGGG	Nanda et al., 2002
Hymenoptera, Formicidae (ants)	TTAGG	Lorite et al., 2002
Paramecium, Tetrahymena	TTGGGG, TTTGGG	McCormick-Graham and Romero, 1996
Plasmodium falciparum	AACCCTA	Gardner et al., 2002
Plasmodium yoelii yoelii	AACCCTG	Carlton et al., 2002

have occurred (Fig. 16.3, step 3) followed by disruption of the *LCN1b* gene in human (but not mouse) (step 4). Finally, the presence of new exons in human *OBPIIa* and *OBPIIb* suggests a selective duplication of individual exons (step 5).

5. **Blocks of Tandemly Repeated Sequences Such As Are Found at Telomeres, Centromeres, and Ribosomal Gene Clusters.** Telomeres are repetitive sequences found at the ends of eukaryotic chromosomes. They provide stability to chromosomes by preventing the degradation of the chromosome end and by blocking the fusion of chromosome ends. Several telomere repeat sequences are listed in Table 16.6. In human telomeres, the short sequence TTAGGG is repeated thousands of times. Try a blastn search using TTAGGG TTAGGG TTAGGG TTAGGG as a query, restricting the output to human, and remove the filter for low complexity. The result is several thousand BLAST hits, most from telomeric sequences such as that shown in Figure 16.4.

The centromere is a constricted site of a chromosome that serves as an attachment point for spindle microtubules, allowing chromosomal segregation during mitotic and meiotic cell divisions (Choo, 2001). All eukaryotic chromosomes have a functional centromere, although the primary nucleotide sequence is not well conserved between species. In humans, this DNA consists largely of a 171-bp repeat of

```
>gi|7407196|gb|AF236885.1|AF236885 Homo sapiens clone p10 chromosome 6,
telomeric repeat region
GGATCCCCCCCAACTCATGACTGTCGGGCTATTTCCAGGCCGCATCGACAGTGAACAAAATCCTTTCTGT
TTGCAGCCCTGAATAATCAGGGTTAGGGTTAGGGTTAGGGGTTAGGGGTTGGGGTTGGGGTTAGGGTTAG
GGTTGGGTTAGGGTTAGGGTTAGGGGTTAGGGTTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTAGGGTC
AGGGTCAGGGTCAGGGTCAGGGTCAGGGTTAAGGGTCAGGGTCAGGGTCAGGGTCAGGGTCAGGGGTTAG
GGTTAGGGTTAGGGTCAGGGTCAGGGTCAGGGTCAGGGTCAGGGGTTAGGGTCGGGGTCGGGGTCGGGGT
CAGGGGTCAAGGGTCAAGGGTCGGGGTCAGGGGTCAAGGGTCGGGGTCGGGGTCCGGGTCAGGGTGAGGG
TGAGGGTGAGGGTGAGGGTGGGGG
```

FIGURE 16.4. *As an example of eukaryotic telomeric repeats, a blastn search of human DNA with the query $(TTAGGG)_4$ results in dozens of matches, including this clone (GenBank accession AF236885). TTAGGG repeats are highlighted in red and underlined, while related TCAGGG repeats are indicated with a dotted underline. Thousands of such repeats occupy the telomeres.*

Lipocalins localized to human chromsome 9q32-34 include α-1-microglobulin/bikunin (NM_001633), complement component 8, gamma polypeptide (NM_000606), lipocalin 1 (protein migrating faster than albumin, tear prealbumin; *LCN1*) (NM_002297), lipocalin 2 (oncogene 24p3)(NM_005564), odorant-binding protein (OBP) 2A (NM_014582) and 2B (NM_014581), orosomucoid 1 (NM_000607) and 2 (NM_000608), progestagen-associated endometrial protein (NM_002571), and prostaglandin D2 synthase (NM_000954). See Chan et al. (1994) and Dewald et al. (1996).

The function of *LCN1* is not known; it was identified by cloning a cDNA from a tear gland library. Rat and bovine OBPs selectively bind odorants of many diverse chemical classes (e.g., terpenes, aldehydes, esters, and musks) (Pevsner et al., 1990; Pelosi, 1996). Thus it is assumed that human OBP gene products also transport hydrophobic ligands.

The Mouse Genome Sequencing Consortium (2002) described a group of eight lipocalin genes on the mouse X chromosome that are absent from primates (see page 1020–44 below). These may have been generated by local gene duplication.

Telomeric repeats are synthesized by telomerase, a ribonucleoprotein that has specialized reverse transcriptase activity. A telomere database (TelDB) is available online at ▸http://www.genlink.wustl.edu/teldb/index.html.

The GenBank accession number for a human α-satellite consensus sequence is X07685. A blastn seatch using this as a query results in over 20,000 database hits. If you exclude human entries from the output of your search (with the command "satellite NOT human[organism]"), you will find that the human α-satellite sequence matches other primates. However, the human sequence is not homologous to nonprimate sequences.

α-satellite DNA extending for 1 to 4 Mb. Almost all eukaryotic centromeres are able to bind a histone H3-related protein (called CENP-A in vertebrates). This protein-DNA complex forms a building block of centromeric chromatin that is essential for the function of the kinetochore, the site of attachment of the spindle fiber.

Satellite DNA is a feature of every known eukaryotic centromere, with only two documented exceptions. In the yeast *S. cerevisiae*, the entire centromere sequence extends only several hundred base pairs. A second exception is the neocentromere, an ectopic centromere that assembles a functional kinetochore, is stable in mitosis, but lacks α-satellite DNA (Amor and Choo, 2002). About 60 human neocentromeres have been described, many involving trisomy or tetrasomy (extra chromosomal copies).

Software to Detect Repetitive DNA: RepeatMasker and Censor

RepBase Update has been developed since 1990 by Jerzy Jurka and colleagues. RepeatMasker was written by Arian Smit and Phil Green. It is available at ▸http://www.geospiza.com/products/tools/repeatmasker.htm.

Finding repetitive DNA elements in eukaryotic DNA, such as LINEs and SINEs, is essential in genome analysis. Knowing the location of repeats can be helpful in identifying likely noncoding regions, and the repeats themselves can be "molecular fossils" that are useful for the comparative analysis of genomes from different species (Chapter 17). A practical way to find repetitive DNA in genomic sequence is to search against a database of known repeats and low-complexity regions. RepBase Update is the premier database of repetitive elements in eukaryotes. Several programs, including RepeatMasker and the Censor Server at GIRI, effectively allow searches of DNA query sequences against this database (Smit, 1999; Jurka, 2000).

The Censor Server of the Genetic Information Research Institute (GIRI) is at ▸http://www.girinst.org/.

The genomic DNA we are using is from contig NT_008769.

To identify and mask repetitive DNA sequences, you can use a RepeatMasker web server. Several servers are listed in Table 16.7, making it unnecessary to install either the program or the database locally. We will explore this with 100,000 bp of genomic DNA from human chromosome 10, a region that includes the *RBP4* gene. Paste your sequence into a box that is provided and select the output options (Fig. 16.5). The GIRI censor server offers a similar web interface (Fig. 16.6). The RepeatMasker output includes a list of scores using the Smith-Waterman algorithm,

TABLE 16-7 Web Servers That Provide Access to Software for Identifying Repetitive Elements in Genomic DNA

Program	Description	URL
RepeatFinder	A computational system for analysis of repetitive structure of genomic sequences	▸http://www.tigr.org/softlab/
RepeatMasker	University of Washington Genome Center	▸http://repeatmasker.genome.washington.edu
RepeatMasker	Server at EMBL	▸http://woody.embl-heidelberg.de/repeatmask/
RepeatMasker	Server in Barcelona	▸http://humangen.med.ub.es/tools/RepeatMasker.html
RepeatMasker	For zebrafish; at the Wellcome Trust Sanger Instiutte	▸http://www.sanger.ac.uk/Projects/D_rerio/fishmask.shtml
Censor Server	Genetic Information Research Institute	▸http://www.girinst.org/

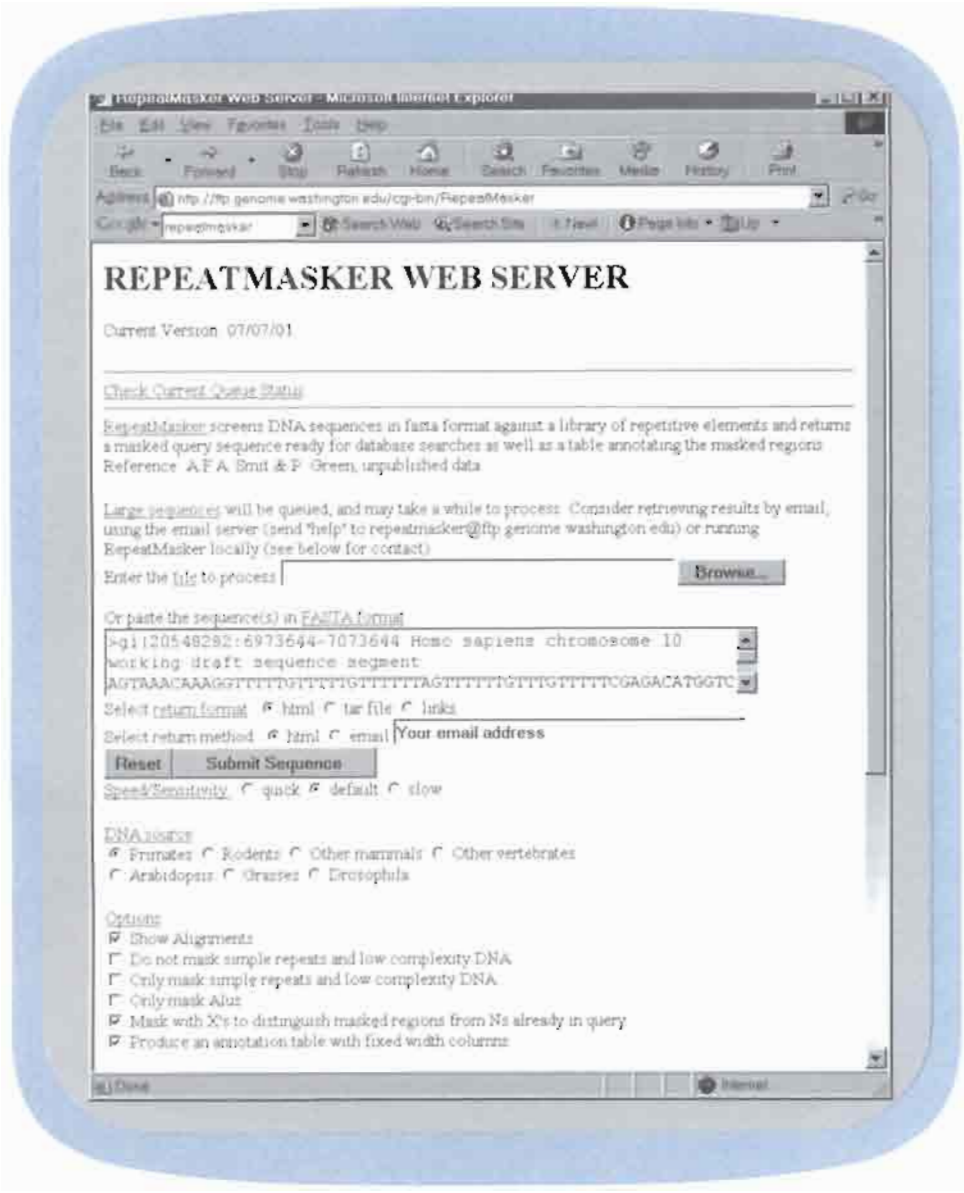


FIGURE 16.5. *The RepeatMasker web server allows repetitive elements in genomic DNA to be identified. The input is 100,000 bp of human genomic DNA from chromosome 10.*

the position of the repeat, and information on the type of repeat (e.g., SINE/*Alu*, LTR, or simple repeat) (Fig. 16.7). Pairwise alignments are displayed (Fig. 16.8). The input sequence is also returned in the FASTA format, with the repetitive residues masked with the letter N or X (Fig. 16.9). This version of the sequence is especially useful for subsequent database searches. Finally, the RepeatMasker program provides a summary of the repetitive elements that were identified (Fig. 16.10).

Recall that a BLAST search uses the SEG and/or DUST programs to define and mask repetitive DNA sequences and also to detect and mask low-complexity protein sequences (Chapter 4).

Definition of Gene

We have begun our analysis of eukaryotic genomes by considering noncoding and repetitive DNA. The coding portions of a genome are of particular interest, as they largely determine the phenotype of all organisms. Two of the biggest challenges in understanding any eukaryotic genome are defining what a gene is and identifying

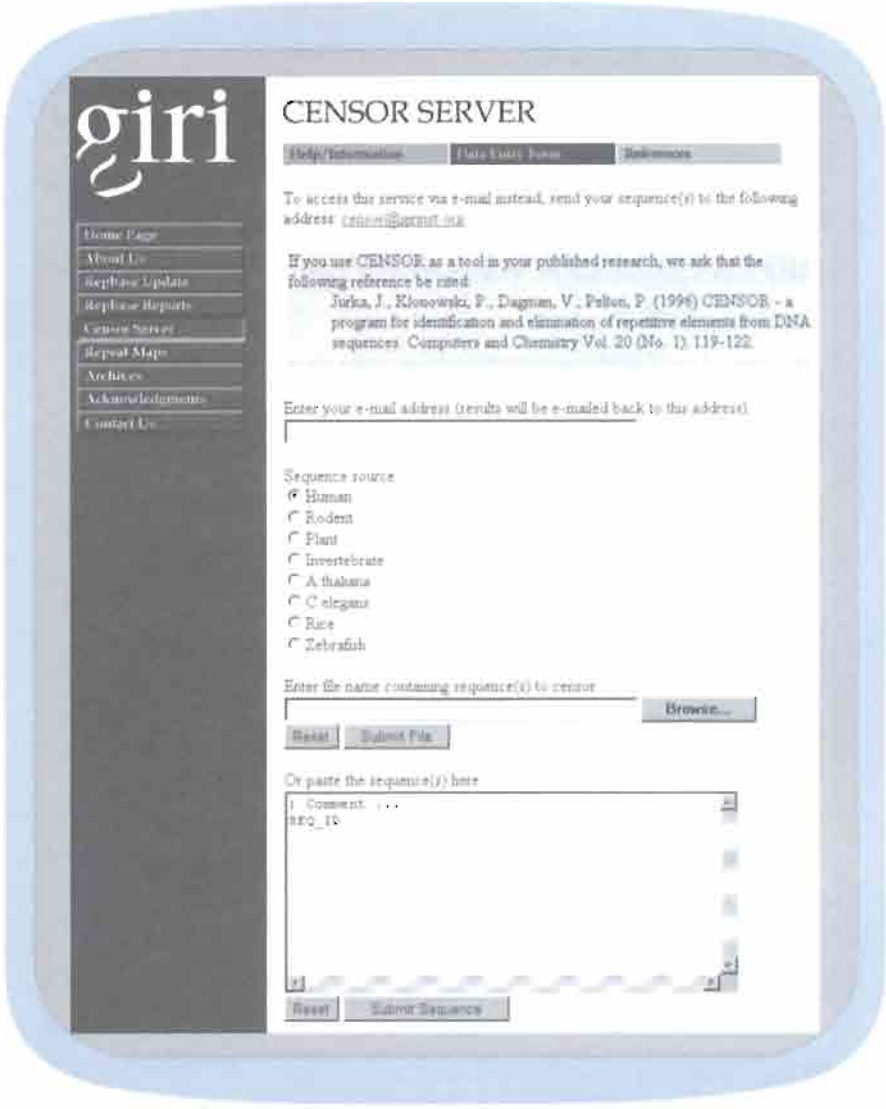


FIGURE 16.6. The GIRI Censor Server, like RepeatMasker, identifies repetitive elements in genomic DNA.

genes within genomic DNA. We will first define the variety of genes and then give the criteria for identifying them:

- Protein-coding genes form a major category of genes. Several criteria are applied to the assignment of a DNA sequence as a protein-coding gene. The principal requirement is that there must be an open reading frame (ORF) of at least some minimum length such as 90 bp (corresponding to 30 codons encoding amino acids, or an ≈3-kD protein).
- Pseudogenes, described above, do not encode functional gene products.
- In addition to pseudogenes, many other kinds of genes also do not encode protein, but instead encode functional RNA molecules (Eddy, 2001, 2002). These include transfer RNA (tRNA) genes. These translate information from the triplet codons in mRNA to amino acids. One program, tRNAscan-SE, identifies 99–100% of RNA genes in genomic DNA sequence with an error rate of one false positive per 15 Gb (Lowe and Eddy, 1997). An example of the tRNAscan-SE server is shown in Figure 16.11.

The tRNAscan-SE search server is at ▸http://www.genetics.wustl.edu/eddy/tRNAscan-SE/. A related resource is the genomic tRNA database (GtRDB), available at ▸http://rna.wustl.edu/GtRDB/. This database summarizes the types and quantities of tRNA molecules from several eukaryotes as well as dozens of bacteria and archaea.

```
RepeatMasker completed 05-Nov-2002 12:15:50 PST
Repeat sequence:
   SW  perc perc perc    position in query      matching       repeat          position in  repeat
score  div. del. ins.    begin    end  (left)    repeat         class/family     begin  end (left)  ID

 1446  13.8  2.8 10.4       19    223 (99287) C  AluJo     SINE/Alu          (1)   311    137   1
 2438   7.3  0.3  0.3      224    525 (98985) C  AluYa5    SINE/Alu          (9)   302      1   2
 1446  13.8  2.8 10.4      526    637 (98873) C  AluJo     SINE/Alu        (175)   137     18   1
  823  14.8  0.0  2.3     1025   1152 (98358) C  FLAM_C    SINE/Alu         (18)   125      1   3
  251  32.5  3.6  8.3     1201   1361 (98149) +  MIR       SINE/MIR           9   173   (89)   4
 2180  13.5  0.7  0.0     1362   1665 (97845) +  AluSq     SINE/Alu           1   306    (7)   6
  251  32.5  3.6  8.3     1666   1749 (97761) +  MIR       SINE/MIR         173   259    (3)   4 *
  684  30.3  5.2  0.9     1690   1920 (97590) C  MIR       SINE/MIR        (16)   246      6   7
  392  22.2  0.0  1.3     2514   2612 (96898) C  MLT1I     LTR/MaLR         (0)   450    319   8
 2335  10.1  0.0  2.8     2705   3022 (96488) C  AluSq     SINE/Alu         (4)   309      1  10
  380  19.4 14.7  2.3     3033   3161 (96349) C  MLT1J2    LTR/MaLR       (272)   178     34  11
  314  26.1  9.2  2.5     3354   3472 (96038) +  MER34B    LTR/ERV1           5   131   (434)  12
  186  27.0  0.0  0.0     3474   3536 (95974) +  (TGGG)n   Simple_repeat      2    64    (0)  13 *
  588  24.4  0.0  0.0     3530   3709 (95801) +  (TGGA)n   Simple_repeat      1   180    (0)  14
  215  26.5  0.0  0.0     3710   3758 (95752) +  (TGGG)n   Simple_repeat      4    52    (0)  15
  363  20.2  5.0  8.5     3871   3956 (95554) +  MER34C    LTR/ERV1         320   407  (168)  12
 2026  14.8  2.1  0.0     3957   4246 (95264) C  AluJb     SINE/Alu        (14)   298      3  17
  363  20.2  5.0  8.5     4247   4384 (95126) +  MER34C    LTR/ERV1         407   544   (31)  12
 2161  10.3  1.0  0.3     4896   5186 (94324) +  AluSp     SINE/Alu           1   293   (20)  20
  337  10.6  0.0  0.0     5355   5428 (94082) C  Alu       SINE/Alu         (0)   296    223  22
  248   6.8 11.4  0.0     5423   5466 (94044) +  MADE1     DNA/Mariner       31    79    (1)  23 *
  386  24.1  7.5  1.5     5474   5606 (93904) C  MLT1F     LTR/MaLR         (0)   542    402  24
  231  16.7  0.0  6.2     5624   5671 (93839) C  MLT1F2    LTR/MaLR       (389)   206    162  24
 2134   9.7  0.0  4.4     5674   6002 (93508) C  AluSp     SINE/Alu         (0)   316      1  27
 2046  10.7  0.0  0.0     6003   6272 (93238) C  AluSq     SINE/Alu        (13)   300     31  28
  320  29.3  4.1  1.6     6281   6403 (93107) C  MLT1F2    LTR/MaLR       (436)   126      1  24
  221  36.2  9.6  0.0     6555   6692 (92818) C  MIR       SINE/MIR        (66)   188     38  30
  233  21.9 12.5  0.0     6912   6975 (92535) +  L1ME4a    LINE/L1         5530  5601  (520)  34
  213  21.1  0.0  0.0     7187   7224 (92286) +  (CA)n     Simple_repeat      1    38    (0)  33
  459  25.1  7.2  6.0     7335   7566 (91944) +  L1ME4a    LINE/L1         5791  6030   (91)  34
 2413   9.2  0.0  0.3     7567   7872 (91638) C  AluSg     SINE/Alu         (5)   305      1  35
  459  25.1  7.2  6.0     7873   7958 (91552) +  L1ME4a    LINE/L1         6030  6113    (8)  34
  215  29.7  1.8  4.4     8068   8240 (91270) C  MIR       SINE/MIR        (27)   235     41  36
  443  26.9  4.9  5.8     8496   8718 (90792) +  MLT1K     LTR/MaLR        310   530   (61)  38
```

FIGURE 16.7. *RepeatMasker output shows the scores and positions of a variety of repetitive DNA elements such as* Alu *repeats and LINE and SINE elements.*

- Ribosomal RNA (rRNA) genes also function in translation.
- Small nucleolar RNAs (snoRNAs) function in the nucleolus.
- Small nuclear RNAs function in the spliceosomes that remove introns from primary RNA transcripts.
- MicroRNAs (miRNAs) are about 21–25 nucleotides in length and are widely conserved among species and may serve as antisense regulators of other RNAs (Ambros, 2001; Ruvkun, 2001).

In annotating genomic DNA, an emphasis is often placed on describing the protein-coding genes. However, it is now clear that noncoding genes encoding various types of RNA products have diverse and important functions. Furthermore, it is not as straightforward to identify noncoding RNAs (Eddy, 2002). Their full size might be extremely small, as in the case of miRNAs. There is no ORF to help define the boundaries of noncoding genes. Database searches may be less sensitive than is possible for protein-coding genes, because the scoring matrices for amino acids are more sensitive and specific. Several databases of noncoding RNAs are listed in Table 16.8 such as Rfam (Griffiths-Jones et al., 2003), RTNBase (Murthy and Rose, 2003), and the noncoding regulatory RNA database (Szymański et al., 2003).

Finding Genes in Eukaryotic Genomes

Finding protein-coding genes in eukaryotic genomes is a far more complex problem than for prokaryotes (Burge and Karlin, 1998; Mural, 1999; Claverie, 1997).

```
186 26.98 0.00 0.00 gi|20548282:6973644-7073644 3474 3536 (95974)
C (CCCA)n#Simple_repeat (116) 64 2 * 5

  gi|20548282:6     3474 TGGATGTGTGGGTGAATGGGCAGCTGGATGGATGAGTGGGCAGGTAGATA 3523
                            i  v       ii    ii v   i   i  i     ii   i i i
C (CCCA)n#Simple_     64 TGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTGGGTG 15

  gi|20548282:6     3524 AGTGGGTGGATGG 3536
                         i         i
C (CCCA)n#Simple_     14 GGTGGGTGGGTGG 2

Transitions / transversions = 7.50 (15 / 2)
Gap_init rate = 0.00 (0 / 63), avg. gap size = 0.00 (0 / 0)

     213 21.05 0.00 0.00 gi|20548282:6973644-7073644 7187 7224 (92286)
     (CA)n#Simple_repeat 1 38 (142) 5

       gi|20548282:6     7187 CACACACACACACTCATGCATGCACACACATATGCACA 7224
                                                v  ii  ii        i ii
       (CA)n#Simple_re      1 CACACACACACACACACACACACACACACACACACACA 38

     Transitions / transversions = 7.00 (7 / 1)
     Gap_init rate = 0.00 (0 / 38), avg. gap size = 0.00 (0 / 0)

2046 10.74 0.00 0.00 gi|20548282:6973644-7073644 6003 6272 (93238)
C AluSq#SINE/Alu (13) 300 31 1

  gi|20548282:6     6003 TTTATTTACTTATTTTTGAGACGGAGTTTCACTTTTGTTTCCCAGACTGG 6052
                           v   vi  v                    i  i     v     i
C AluSq#SINE/Alu     300 TTTTTTTTTTTTTTTTTGAGACGGAGTTTCGCTCTTGTTGCCCAGGCTGG 251

  gi|20548282:6     6053 AGTGCAATGGCGCCATCTTGGCTCAGTGCAACCTCTGCCTCCCAGGTTCA 6102
                            i      v    i      v         i       i
C AluSq#SINE/Alu     250 AGTGCAGTGGCGCGATCTCGGCTCACTGCAACCTCCGCCTCCCGGGTTCA 201

  gi|20548282:6     6103 AGCGATTCTCCTGCTTCAGCCTCCCGAGTAGCTGGGATTACAGGCGCGTG 6152
                                 i                                      vi
C AluSq#SINE/Alu     200 AGCGATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGATTACAGGCGCCCG 151

  gi|20548282:6     6153 CCATCATGCCTGGCTAATTTTTGTATTTTTTGTAGAGACGGGGTTTCACC 6202
                           i  i   i                    v
C AluSq#SINE/Alu     150 CCACCACGCCCGGCTAATTTTTGTATTTTTAGTAGAGACGGGGTTTCACC 101

  gi|20548282:6     6203 ATATCGGCCAGGCTTATCTTGAACTACTGACCTGAGGTGATCCGCCCGCC 6252
                          i i         vi   i     v       v
C AluSq#SINE/Alu     100 ATGTTGGCCAGGCTGGTCTCGAACTCCTGACCTCAGGTGATCCGCCCGCC 51

  gi|20548282:6     6253 TCAGCCTCCCAAAGTGCTGG 6272
                         i
C AluSq#SINE/Alu      50 TCGGCCTCCCAAAGTGCTGG 31

Transitions / transversions = 1.64 (18 / 11)
Gap_init rate = 0.00 (0 / 270), avg. gap size = 0.00 (0 / 0)
```

FIGURE 16.8. *Examples of repetitive sequences identified by RepeatMasker in pairwise alignments. A simple repeat is shown at top, and an* Alu *repeat is shown at bottom.*

While bacterial genes typically correspond to long ORFs, most eukaryotic genes have exons and introns. There are several kinds of exons (Fig. 16.12):

1. Noncoding exons correspond to the untranslated 5′ or 3′ region of DNA.
2. Initial coding exons include the start methionine and continue to the first 5′ splice junction.
3. Internal exons begin with a 3′ splice site and continue to a 5′ splice site.
4. Terminal exons proceed from a 3′ splice site to a termination codon.
5. Single-exon genes are intronless, beginning with a start codon and ending with a stop codon (Table 16.5).

```
Masked Sequence:
>gi|20548282:6973644-7073644 Homo sapiens chromosome 10
working draft sequence segment
AGTAAACAAAGGTTTTTGXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXGCCTCTAAACTAA
AGGTTTTAATGAAGGAGAAGAAAAGCCTTGGGGGAAAAATGCTATTATTC
TGTTTGATTCACAAATTATGCATAATGGCACATGTGCTACCTTGCATGGA
TTATGAAGGCAAGCATTTTCACTTCAGTTTTGTAAGGTAGAGGTAAGGGG
CAGGAGAAGCTGATAATAGAGGATTAAGAAAAAAACTTGTAGAGTATATT
ATTATCAGCATAAACTTAGCAATCTGTTAATTAAATTTTGATCTGTTAAA
TGAGTTTAACAATATGTTGCATATATGCCACAGTAGTAATTTCTTCCCTT
GAAGGAGTGAACTTTACGGAAGTGATTCTGTTTATTGGCACTCAAAAGAG
AAGCACCTCAAATTAAAAAAAATTXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXCAATCAATTTAAATTTTTTCTTTTGTTTTACTGCTTGTTTTACATCAT
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXGTGGCCACTCTTCTATCACCCTAAAGCCAG
AAAATGTATGTGAAAGCACATTGCAGATGGCAAATACTGTCCCAGATTAT
TTTCATTTTTCAGCAATGACTGTAGTGTGGACGGAGCTGGAGAGAATGTG
AGAGAAGAAAGTGACATGACCTGCCCCCAGAACTGCCATTCATTTTACTT
GTCCATTATCACCATCTGTCATCCTCACCTGAAACTAGTATCAACGCAGG
TGCATTCAAATGCAAATAACAGAAAAGACTTACACCATCACGACATTTAT
TGTCTTTTTAAAAAGAAGTATGGAATCAGATGCTCGCAGTTTAGATCAGC
AATTTCACAAGGTTAGGGCTCTGGACTCTCATCTCTATCACTCTGCCCAA
GTTCACATCTCTACTGCTACAGTTCCAAGCCTCACATTCTAATAAGATTT
TTCTTCAAAACCAAGAAGACACCTTGCCACGGTCATGAATGAGCACTCCT
AATATGTTTCTTTTGTTTTATTGGAAAGTCCATTCTTTGGAAATGGAAAT
ATTTATAAGGAAACTCCACAGCACATGTCCCTTACATGTTATTGACTAGA
AATGGATATATATXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

FIGURE 16.9. *RepeatMasker places the symbol X to hide repetitive DNA sequences in the FASTA format. These masked sequences can then be searched against databases to find homologous sequences.*

In addition to the issue of introns, eukaryotic genes also occupy a far smaller proportion of the genome than do prokaryotic genes. Eukaryotic protein-coding genes occupy just 25% of the nematode and insect genomes and less than 3% of the human and mouse genomes.

Algorithms for finding protein-coding genes in eukaryotes can be divided into two categories: homology based (also called extrinsic) and algorithm based (also

```
Summary:
==================================================
file name: RM2sequpload_8327
sequences:           1
total length:     99510 bp  (99510 bp excl N-runs)
GC level:         45.31 %
bases masked:     37952 bp ( 38.14 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:               94         20982 bp    21.09 %
      ALUs           52         14798 bp    14.87 %
      MIRs           42          6184 bp     6.21 %

LINEs:               22          7643 bp     7.68 %
      LINE1           9          4876 bp     4.90 %
      LINE2          11          2508 bp     2.52 %
      L3/CR1          2           259 bp     0.26 %

LTR elements:        18          5980 bp     6.01 %
      MaLRs          11          3316 bp     3.33 %
      ERVL            4           975 bp     0.98 %
      ERV_classI      3          1689 bp     1.70 %
      ERV_classII     0             0 bp     0.00 %

DNA elements:        16          2268 bp     2.28 %
      MER1_type      11          1878 bp     1.89 %
      MER2_type       0             0 bp     0.00 %

Unclassified:        0             0 bp     0.00 %

Total interspersed repeats:    36873 bp    37.05 %


Small RNA:           0             0 bp     0.00 %

Satellites:          0             0 bp     0.00 %
Simple repeats:     16           758 bp     0.76 %
Low complexity:      7           330 bp     0.33 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element


The sequence(s) were assumed to be of primate origin.
RepeatMasker version 07/07/2001 , default mode
run with cross_match version 0.990329
RepBase Update 6.3, vs 05152001
```

FIGURE 16.10. *A RepeatMasker output includes a summary of the types of repetitive elements identified. These results are for a 100-kb fragment of human chromosome 10 in a region including the* RBP4 *gene. Twenty-one percent of the sequence consists of SINEs, and a total of 37% of the sequence consists of interspersed repeats.*

About one-third of all human genes are alternatively spliced. If ESTs are available corresponding to alternatively spliced isoforms, these sequences can be mapped to exons.

called intrinsic) (Stein, 2001). These approaches are outlined in Figure 12.17. Homology-based approaches typically involve the alignment of expressed genes (ESTs from cDNA libraries; see Chapters 2 and 6) with genomic DNA. In these cases, the ESTs can help to define the exon/intron structure in genomic DNA. Thus, homology-based approaches are generally very successful. An additional form of homology-based gene identification is to compare genomic DNA of two related organisms (Morgenstern et al., 2002; Novichkov et al., 2001). By comparing human DNA to pufferfish (*F. rubripes*) DNA, it was possible to discover nearly 1000 putative human genes (Hedges and Kumar, 2002; Aparicio et al., 2002).

While the use of EST data is extremely helpful in annotating eukaryotic genes, there are notable limitations to this approach.

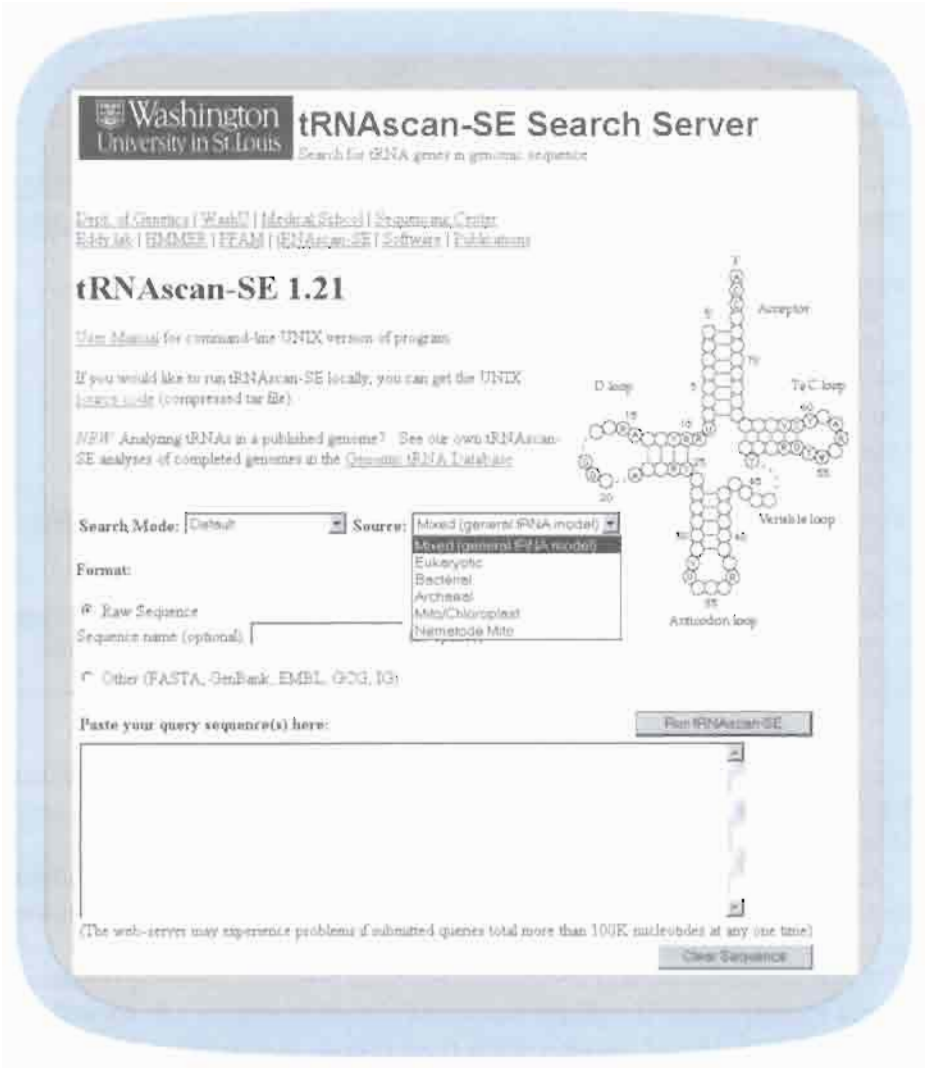


FIGURE 16.11. The tRNAscan-SE search server identifies tRNA molecules in genomic DNA or RNA sequences. This program is available as a web-based version and can be locally installed for larger projects. The tRNAscan-SE program does not contain its own algorithm for finding tRNA molecules, but instead it combines three independent algorithms to find sequence motifs corresponding to conserved tRNA features such as stem-loop structures. tRNAscan-SE predicts the presence of tRNA pseudogenes based on the lack of appropriate secondary-structure motifs.

- The quality of EST sequence is sometimes low, as clones are often sequenced on only one strand and sequencing errors are common.
- Highly expressed genes are often disproportionately represented, although some cDNA libraries are normalized (Chapter 6).
- ESTs provide no information regarding the genomic location.

Intrinsic programs are also widely used to annotate genomic DNA. A large fraction of predicted genes do not have identifiable orthologs, nor are EST sequences available. It is thus essential to identify protein-coding genes using ab initio (intrinsic) approaches. We discussed the GLIMMER program for prokaryotes in Chapter 14.

Many web-based eukaryotic gene prediction programs are available (Table 16.9). These include GENSCAN (Burge and Karlin, 1997) and GRAIL. Rogic et al. (2001) compared seven of these programs in terms of both specificity and sensitivity. The algorithms are improving over time as they account for coding statistics (i.e., the features of nucleotides in coding versus noncoding regions) and signals (e.g., promoter

Kapranov et al. (2002) created DNA microarrays with oligonucleotides corresponding to essentially all of human chromosomes 21 and 22, spaced at 35-nucleotide intervals. Thus these microarrays spanned two chromosomes extremely densely. They hybridized probes derived from a variety of human cell lines and found evidence for the existence of 10 times more RNA transcripts than have been previously annotated. Some of these transcripts could represent novel genes that were not previously annotated while many other possibilities exist—they may have detected a variety of noncoding RNAs or nonfunctional RNA species.

TABLE 16-8 Web-Based Databases of Noncoding RNA

Database	Description	URL
Large-ribosomal-subunit database	Structure of large-subunit ribosomal subunit RNA	▸http://rrna.uia.ac.be/rrna/lsu/
Noncoding RNAs database	Various RNA categories	▸http://biobases.ibch.poznan.pl/ncRNA/
Rfam	RNA family database	▸http://www.sanger.ac.uk/Software/Rfam/ and ▸http://rfam.wustl.edu/
Ribosomal Database Project (RDP)	Provides ribosome-related data analysis, rRNA-derived phylogenetic trees, and aligned and annotated rRNA sequences.	▸http://rdp.cme.msu.edu/html/
RNABase	RNA structures	▸http://www.rnabase.org
RNA Editing website	On all the various types of RNA editing	▸http://www.rna.ucla.edu/
Small-ribosomal-subunit database	A database on the structure of small-subunit ribosomal RNA	▸http://rrna.uia.ac.be/rrna/ssu/
Small RNA database	Small RNAs are broadly defined as the RNAs not directly involved in protein synthesis.	▸http://mbcr.bcm.tmc.edu/smallRNA/smallrna.html
tRNA sequences	Compilation of 550 sequences of tRNAs and 3704 sequences of tRNA genes (through 1998).	▸http://www.uni-bayreuth.de/departments/biochemie/sprinzl/trna/index.html

The results of the Rogic et al. (2001) study are posted on the web at ▸http://www.cs.ubc.ca/~rogic/evaluation/.

A simple tool for predicting the presence of protein-coding sequences is the ORF finder from NCBI. Input the genomic DNA sequence of HIV-1, and the program predicts a series of ORFs. Comparison of these ORFs to the authentic HIV-1 coding sequences reveals that there are many putative ORFs that are not authentically transcribed.

elements, start and stop codons, splice sites, and polyadenylation sites). Makarov (2002) also compared seven commonly used programs for eukaryotic gene prediction with an emphasis on availability as well as prediction accuracy.

Although the algorithms used for various gene-finding programs vary slightly, they all suffer from the same general problems, a high false-positive rate combined with a failure to accurately predict complete gene structures.

To see how a eukaryotic gene-finding algorithm performs, try using the Genome Analysis Pipeline at the Oak Ridge National Laboratory (Fig. 16.13). We will again use about 100,000 bp of genomic DNA from human chromosome 10. Select the

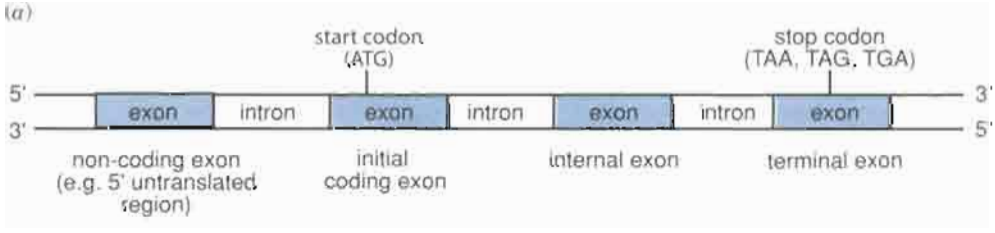


FIGURE 16.12. (a) *Eukaryotic gene prediction algorithms differentiate several kinds of exons, including those in noncoding regions; initial coding exons that include a start codon; internal exons; and terminal exons that include a stop codon. These exons are built into a model for a predicted gene.* (b) *In some cases, genes have a single exon and are intronless. The border of exons and introns typically has a GT/AG boundary, but the structure of genes is still difficult to predict ab initio.*

TABLE 16-9 Algorithms for Finding Genes in Eukaryotic DNA

A summary of sites is given at ▸http://linkage.rockefeller.edu/wli/gene/.

Program	Description	URL
AAT	Analysis and Automation Tool; web-based server	▸http://genome.cs.mtu.edu/aat.html
FGeneH	Predicts exons using linear discriminant functions	▸http://genomic.sanger.ac.uk/gf/gf.html
FgeneSH	Ab initio gene finder	▸http://www.softberry.com/berry.phtml
Gene Finder	For human, mouse, *Arabidopsis*, and fission yeast	▸http://argon.cshl.org/genefinder/
GeneParser2	Identification of protein-coding regions in genomic DNA	▸http://beagle.colorado.edu/%7Eeesnyder/GeneParser.html
Genie	Based on HMMs	▸http://www.cse.ucsc.edu/~dkulp/cgi-bin/genie
GenLang	Syntactic pattern recognition system; uses computational linguistics to find genes	▸http://www.cbil.upenn.edu/genlang/genlang_home.html
Genscan	Based on HMMs; rule based rather than homology based	▸http://genes.mit.edu/GENSCAN.html
GenTerpret	From RabbitHutch Biotechnology Corporation	▸http://www.rabbithutch.com/
GlimmerM	From TIGR	▸http://www.tigr.org/software/glimmerm/
GlimmerM web server	For *Arabidopsis thaliana*, *Oryza sativa* (rice), *Plasmodium falciparum* (malaria), other	▸http://www.tigr.org/tdb/glimmerm/glmr_form.html
GRAIL	One of the most widely used algorithms	▸http://compbio.ornl.gov/tools/index.html
MORGAN	A decision tree system for finding genes in vertebrate DNA	▸http://www.tigr.org/~salzberg/morgan.html
PROCRUSTES	Gene Recognition via Spliced Alignment	▸http://www-hto.usc.edu/software/procrustes/index.html
VEIL	HMM for finding genes in vertebrate DNA	▸http://www.cs.jhu.edu/labs/compbio/veil.html
Xpound	A probabilistic model for detecting coding regions	▸http://bioweb.pasteur.fr/seqanal/interfaces/xpound-simple.html

Abbreviation: HMM, hidden Markov model.

appropriate organism and algorithms offered at the website, and the output includes the following (Fig. 16.14):

- Gene-finding predictions of the GrailEXP program, which identified five genes, then automatically searched a protein database with blastp as well as Pfam.
- Gene-finding predictions from Genscan, which identified three genes.
- Identification of CpG islands, which are potentially associated with housekeeping genes (see Chapter 17).
- RepeatMasker repeats, of which 201 were found (compare Fig. 16.10).
- Transfer RNA scan predictions of tRNA genes (of which none were found).
- Grail bacterial artificial chromosome (BAC) pairs, of which 24 were identified.

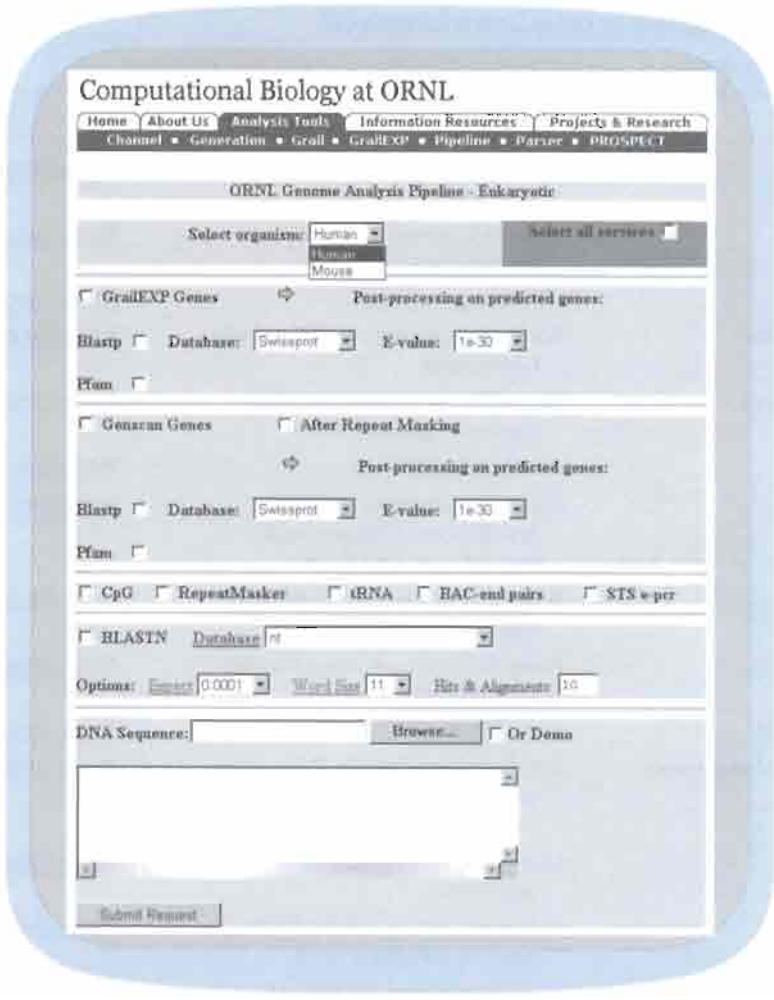


FIGURE 16.13. The Oak Ridge National Laboratory (ORNL) Genome Analysis Pipeline is a web-based tool for the annotation of genomic DNA from several species. It includes the GrailEXP program. This program used 100,001 bp of genomic DNA from human chromosome 10 as input, including RBP4. *The sequence was obtained from NT_008769 (10,921,016 bp of genomic DNA). This pipeline is available at* ▸http://compbio.ornl.gov/tools/pipeline/.

- E-PCR STSs, that is, electronic polymerase chain reaction (PCR) sequence tag sites. These are short DNA sequences of confirmed genomic location that are offered as a service from the E-PCR site on the NCBI home page (▸http://www.ncbi.nlm.nih.gov).

Thus, this analysis pipeline offers a wealth of data, including gene predictions based on independent approaches. The inconsistencies in the actual gene predictions highlight the inherent difficulties in intrinsic approaches to gene prediction. A GrailEXP tool, Perceval, predicted several initial, internal, and terminal exons in the 100,000-bp segment we searched (Fig. 16.15). The estimates of the quality of these predictions varied from marginal to good to excellent. For the five genes predicted by GrailEXP (Fig. 16.16), we can evaluate their accuracy simply by BLAST searching the predicted protein products. From this we learn that the genes identified as genes 3 and 4 actually match different exons of the *RBP4* gene (Fig. 16.17), and the program failed to assemble a correct gene model. This can be seen by comparing the predicted results to those based on experimental evidence (Fig. 16.18). Predicted genes 1, 2, and 5 match other genes that are adjacent to *RBP4* on chromosome 10.

The difficulty of finding protein-coding genes in genomic DNA is illustrated by the efforts to annotate a typical eukaryotic genome: the *indica* and *japonica* subspecies of the rice genome. Yu et al. (2002) obtained 75,659 gene predictions when

Grail Experimental Gene Discovery Suite is available on the web at ▸http://compbio.ornl.gov/grailexp/

For another example of the difficulty in annotating eukaryotic genomic DNA for protein-coding genes, see the discussion of the *Drosophila* genome below (page 590).

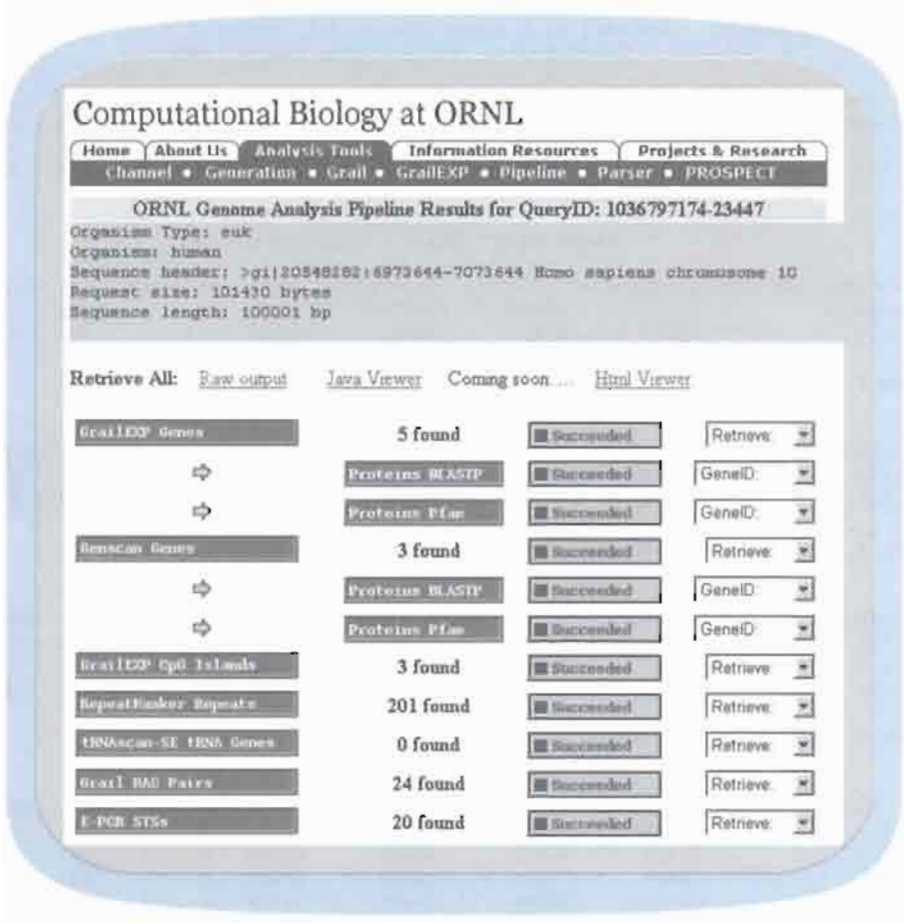


FIGURE 16.14. The ORNL Genome Analysis Pipeline of a human genomic DNA fragment reported five genes using GrailEXP and three genes using GenScan. The pull-down menus at right provide access to Blastp results with the predicted proteins and Pfam alignments. These algorithms correctly identified RBP4 in the genomic DNA. The output also includes RepeatMasker repeats, tRNA genes (although none were identifed in this query), identifiers for BAC clones that contain this DNA region, and sequence-tagged sites (STSs) from the NCBI electronic PCR (e-PCR) site that confirm the assignment of this DNA to chromosome 10.

they submitted their assembled draft version of the rice genome (*indica*) to an FGeneSH web server (see Table 16.9). Only 53,398 of these predictions were complete (having both initial and terminal exons): About 7500 had only an initial exon, 11,000 had only a terminal exon, and 3400 predicted genes had neither. Additionally, they reported that exon–intron boundaries were often not precisely defined. However, when the finished sequence was obtained rather than the draft sequence,

PERCEVAL Exon Candidates (36 predicted)

Index	Std	Begin	End	Frm	Type	Len	Scr	Quality
10	+	35515	35557	0	Internal	43	61	Marginal
11	-	40500	40616	2	Internal	117	81	Good
12	+	43576	43965	0	Terminal	390	92	Excellent
13	-	50105	50391	1	Terminal	287	99	Excellent
14	-	52382	52436	2	Internal	55	53	Marginal
15	-	53575	53701	0	Internal	127	88	Good
16	-	54457	54518	0	Initial	62	44	Marginal
17	-	56741	56855	2	Terminal	115	90	Excellent
18	-	57023	57384	0	Initial	362	81	Good
19	-	57274	57384	0	Initial	111	85	Good
20	+	63682	63812	0	Internal	131	100	Excellent
21	+	69066	69462	2	Internal	397	88	Good
22	+	76988	77140	0	Internal	153	91	Excellent
23	+	77247	77336	0	Internal	90	90	Excellent
24	+	78288	78428	0	Internal	141	94	Excellent
25	+	81931	82005	0	Internal	75	93	Excellent
26	+	82996	83052	0	Internal	57	91	Excellent

FIGURE 16.15. The PERCEVAL program of GrailEXP predicts the existence of exons, their type (initial, internal, or terminal), their length and score, and the quality of the prediction. For 100,000 bp from human chromosome 10, the program predicted 36 exons, some of which are shown here.

Gene 1, Variant 1 Strand: + Bounds: 18563-23742 Exons: 2
Start Codon: Yes Stop Codon: Yes

---Index----	--------Exons--------		---------CDS---------		-Ph-	-Fr-	-Len-	-Scr-
Promoter	14237	14350	...	...	.	.	114	51
1.1.1	18563	18775	18563	18775	0	1	213	70
1.1.2	23068	23742	23068	23742	0	0	675	88

Gene 2, Variant 1 Strand: + Bounds: 43687-43965 Exons: 1
Start Codon: Yes Stop Codon: Yes

---Index----	--------Exons--------		---------CDS---------		-Ph-	-Fr-	-Len-	-Scr-
Promoter	39126	39349	...	...	.	.	224	70
2.1.1	43687	43965	43687	43965	0	0	279	92

Gene 3, Variant 1 Strand: - Bounds: 50105-50413 Exons: 1
Start Codon: Yes Stop Codon: Yes

---Index----	--------Exons--------		---------CDS---------		-Ph-	-Fr-	-Len-	-Scr-
PolyA	45824	45829	...	...	.	.	6	65
3.1.1	50105	50413	50105	50413	0	1	309	97

Gene 4, Variant 1 Strand: - Bounds: 56741-57384 Exons: 3
Start Codon: Yes Stop Codon: Yes

---Index----	--------Exons--------		---------CDS---------		-Ph-	-Fr-	-Len-	-Scr-
4.1.3	56741	56855	56741	56855	2	1	115	90
4.1.2	57023	57159	57023	57159	0	2	137	71
4.1.1	57274	57384	57274	57384	0	2	111	85

Gene 5, Variant 1 Strand: + Bounds: 63682-97389 Exons: 14
Start Codon: No Stop Codon: Yes

---Index----	--------Exons--------		---------CDS---------		-Ph-	-Fr-	-Len-	-Scr-
5.1.1	63682	63812	63682	63812	0	0	131	100
5.1.2	69066	69462	69066	69462	2	0	397	88
5.1.3	76988	77140	76988	77140	0	1	153	91
5.1.4	77247	77336	77247	77336	0	2	90	90
5.1.5	78288	78428	78288	78428	0	2	141	94
5.1.6	81931	82005	81931	82005	0	0	75	93
5.1.7	82996	83052	82996	83052	0	0	57	91
5.1.8	85614	85661	85614	85661	0	2	48	65
5.1.9	91114	91267	91114	91267	0	0	154	100
5.1.10	91920	91996	91920	91996	1	1	77	84
5.1.11	93351	93419	93351	93419	0	2	69	97
5.1.12	96426	96572	96426	96572	0	2	147	93
5.1.13	96806	96913	96806	96913	0	1	108	93
5.1.14	97276	97389	97276	97389	0	0	114	100
PolyA	98736	98741	...	...	.	.	6	79

FIGURE 16.16. *GrailEXP predicted the existence of five genes. Predicted genes 3 and 4 form part of the authentic* RBP4 *gene (see Fig. 16.18).*

the estimate of gene content improved dramatically. Sasaki et al. (2002) obtained the finished sequence of rice chromosome 1 (subspecies *japonica*) and predicted 6756 genes on this chromosome. In contrast, the draft version of this genome predicted just 4467 genes. Sasaki et al. (2002) suggest that the presence of several thousand gaps in the draft sequence precluded the ability to accurately predict complete genes.

Protein-Coding Genes in Eukaryotes: New Paradox

The *C* value paradox is answered based on the variable amounts of noncoding DNA in a variety of eukaryotes. A new paradox is introduced: Why are the proteomes of various eukaryotes similar in size, given the enormous phenotypic differences between eukaryotes? Claverie (2001) calls this the *N* value paradox (*N* is for number), while Betrán and Long (2002) call this the *G* value paradox (*G* is for genes). As we survey eukaryotic genomes in the second half of this chapter, we will see that

Gene 3 (predicted by GrailEXP) as query for a blastp nr search

```
>gi|5803139|ref|NP_006735.1|   retinol-binding protein 4, plasma precursor;
          retinol-binding protein 4, plasma; retinol-binding protein 4,
          interstitial [Homo sapiens]
 gi|72085|pir||VAHU   plasma retinol-binding protein precursor - human
 gi|35897|emb|CAA24959.1|   precursor RBP [Homo sapiens]
         Length = 199

 Score =  152 bits (385), Expect = 5e-38
 Identities = 71/72 (98%), Positives = 72/72 (99%)

Query: 8   DDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELC 67
           +DDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELC
Sbjct: 117 NDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELC 176

Query: 68  LARQYRLIVHNG 79
           LARQYRLIVHNG
Sbjct: 177 LARQYRLIVHNG 188
```

Gene 4 (predicted by GrailEXP) as query for a blastp nr search

```
 Score =  204 bits (518), Expect = 2e-53
 Identities = 105/120 (87%), Positives = 107/120 (88%), Gaps = 2/120 (1%)

Query: 1   MKWVWXXXXXXXXGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV 60
           MKWV              AERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
Sbjct: 1   MKWV--WALLLLAAWAAAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV 58

Query: 61  AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGSE 120
           AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKG+→
Sbjct: 59  AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND 118
```

FIGURE 16.17. *The GrailEXP output includes predicted proteins, provided in the FASTA format. When these are searched by blastp against the nonredundant (nr) database, the gene 3 and gene 4 protein products can be seen to match* RBP4. *For gene 3, the first eight predicted amino acids do not match authentic* RBP4.

organisms such as worms and flies appear to have about 13,000–20,000 protein-coding genes, while plants, fish, mice, and humans have only slightly more (about 25,000–40,000 genes) (Harrison et al., 2002). Why do organisms such as humans, having so much greater biological complexity than insects and nematodes, have not quite twice as many genes? The genes of higher eukaryotes may employ more complex forms of gene regulation, such as alternative splicing.

Transcription Factor Databases and Other Genomic DNA Databases

In addition to predicting the presence of genes, it is also important to predict the presence of genomic DNA features such as promoter elements (e.g., transcription factor binding sites). Algorithms are available for these tasks, as well as databases storing compilations of genomic features (Table 16.10).

A database of introns from many organisms is available from ▸http://www.introns.com/.

A codon usage database is available at ▸http://www.kazusa.or.jp/codon/.

Several gene-finding algorithms report GC content as part of their output (Knight et al., 2001). Gentles and Karlin (2001) compared the dinucleotide

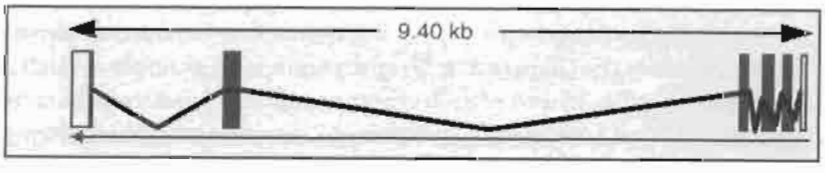


FIGURE 16.18. *The actual structure of the* RBP4 *gene (from the Ensembl web server; see Chapter 17) includes six exons, according to the Ensembl gene report and according to the evidence viewer at NCBI (available via the map viewer or LocusLink). It is extremely difficult for algorithms such as GrailEXP to find all the exons and assemble them into a gene model in the absence of experimental evidence such as the full-length cDNA (or protein) sequence.*

TABLE 16-10 Software for Identifying Features of Promoter Regions in Genomic DNA

Program	Description	URL
Ancient conserved untranslated DNA sequences (ACUTS)	Analyzes genes from metazoan species (essentially vertebrates, insects, and nematods)	▸http://pbil.univ-lyon1.fr/acuts/ACUTS.html
AliBaba2	Predicts binding sites of transcription factor binding sites in an unknown DNA sequence	▸http://www.gene-regulation.de/
Eukaryotic Promoter Database (EPD)	Annotated nonredundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally	▸http://www.epd.isb-sib.ch
FastM	Develops models of transcriptional regulatory DNA units (e.g., promoters).	▸http://www.gene-regulation.de/
Gene Regulation	Requires log-in	▸http://www.gene-regulation.de
PlantProm	Plant promoter database	▸http://mendel.cs.rhul.ac.uk
rSNP_Guide	Transcription factor binding site database	▸http://util.bionet.nsc.ru/databases/rsnp.html
TRANSFAC	Database of transcription factors, their genomic binding sites, and DNA-binding profiles	▸http://transfac.gbf.de

abundances and their biases from human, *S. cerevisiae, A. thaliana, and D. melanogaster*. They found distinctive patterns that are homogeneous within the genome of each species, but markedly different between eukaryotes.

Eukaryotic Genomes Are Organized into Chromosomes

An ideogram is a diagram of a karyotype. A karyotype is an image (often a photograph) of the chromosomes from a cell during metaphase, when each chromosome is a pair of sister *chromatids*. Karyotypes display the chromosomes in numerical order, with the short arm (p arm) oriented upward. The short arm is called "p" for *petit* (French for "small"), while the q arm (long arm) is named as the letter following p.

Chromosomes are often studied at metaphase, when they are thickest and most condensed. Typically, a sample is collected from blood cells or amniotic fluid. Chromosomes are most often visualized using dyes or using specific DNA probes by fluorescence in situ hybridization (FISH).

Genomic DNA is organized in chromosomes. In recent decades it has been possible to use the tools of comparative genomics to describe the relative numbers of genes in species, the relative positions of genes along the chromosomes, and the features of chromosomes such as euchromatin and heterochromatin content, GC content, and structural and regulatory features. With the ability to sequence eukaryotic genomes, it is now possible to examine chromosomes in great detail.

The diploid number of chromosomes is constant in each species. We explored the 16 *S. cerevisiae* chromosomes (Chapter 15), including a variety of databases such as NCBI, MIPS, and SGD that provide graphic displays. In humans, the diploid number is 46 (i.e., there are 23 pairs of chromosomes in all somatic cells). We will explore databases that display ideograms of the human chromosomes in Chapters 17 and 18. Ideograms of karyotypes for some other organisms are available online (Table 16.11).

As we explore a variety of eukaryotic genomes that have been completely sequenced, it is helpful to describe the structure and content of chromosomes. We will refer to a karyotype of human metaphase chromosomes visualized with Wright's stain (Fig. 16.19). A variety of stains produce banding patterns on chromosomes. These include Q bands (based on stains using quinacrine mustard or derivatives) and G bands (based on the Giemsa dye; Wright's stain is an example of such a dye). These dyes stain the entire length of each chromosome and produce a characteristic banding pattern. A band is defined as a portion of a chromosome that is distinguishable from adjacent segments by appearing lighter or darker.

There are several major features of eukaryotic chromosomes. The most apparent landmarks are the telomeres (the chromosome ends) and the centromere, a region that remains unstained with many dyes.

TABLE 16-11 Web-Based Databases of Chromosomes

Resource	Comment	URL
Ensembl genome browser	Ideograms for human (Chapter 17), mouse, rat, zebrafish, fugu, and mosquito	http://www.ensembl.org/
The Unified Database for Human Genome Mapping	From the Weizmann Institute	http://bioinformatics.weizmann.ac.il/cgi-bin/udb/search_map_sbr.pl
Human chromosome-specific databases	From the U.K. HGMP Resource Centre	http://www.hgmp.mrc.ac.uk/GenomeWeb/human-gen-db-chromosomes.html
Ideogram Album	Human, mouse, and horse ideograms from the University of Washington	http://www.pathology.washington.edu/research/cytopages/
Human Chromosome Launchpad	From the Oak Ridge National Laboratory	http://www.ornl.gov/hgmis/launchpad

We might think of chromosomes as unchanging entities that define the genome of each species. However, they are dynamic in many ways across large time scales (millions of years), between generations, and even with individual lifetimes. A broad variety of cytogenetic changes occur in eukaryotes, many of which are disease-causing abnormalities:

- A whole-genome duplication event (autopolyploidy) may occur. Such a massive event happened in yeast (Chapter 15) and plants (see below) and possibly in many other lineages as well.
- The genomes of two distinct species may merge to generate a novel species (allopolyploidy) (Hall et al., 2002). This phenomenon has been described in many plants (Comai, 2000), animals, and fungi. For example, the plant *Arabidopsis suecica* derives from the *A. thaliana* and *Cardaminopsis aerenosa* genomes (Lee and Chen, 2001; Lewis and Pikaard, 2001). Another example of allopolyploidy is the mule, which is the result of a cross between a male donkey (*Equus asinus*, 2n = 62) and a female horse (*Equus caballus*, 2n = 64).

A metacentric chromosome has its centromere located near the center of the chromosome. An acrocentric chromosome has a centromere near a telomere. In mouse, all the chromosomes are acrocentric. For humans, the acrocentric chromosomes are 13, 14, 15, 21, and 22. These acrocentric chromosomes may be subject to Robertsonian translocation, in which two centromeres fuse (Slijepcevic, 1998a).

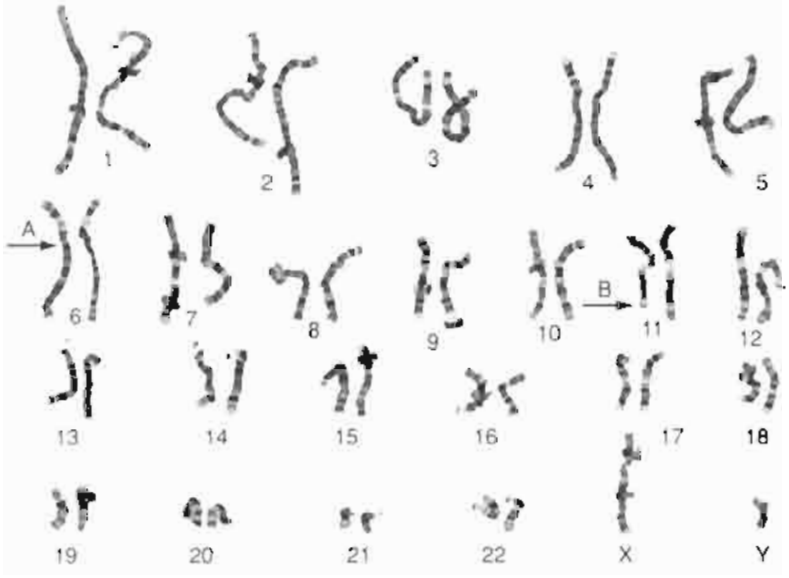


FIGURE 16.19. Example of a human karyotype. The chromosomes are visualized with Wright's stain. Centromeres are visible as an indentation in the chromosome (e.g., see arrow A). This karyotype is of a person with a hemizygous deletion of a telomeric portion of chromosome 11q (arrow B).

Mules cannot propagate because they are sterile (they cannot produce functional halploid gametes) (see Ohno, 1970).

- An individual may acquire an extra copy of an entire chromosome. For example, Down syndrome is caused by a trisomy (triplicated copy) of chromosome 21. We will discuss this type of disorder in Chapter 18.
- Uniparental disomy may occur, in which both homologous chromosomes are inherited from one parent. This is also associated with disease in humans (Kotzot, 2001).
- Chromosomes can fuse. Human chromosome 2, the second largest human chromosome, is derived from two ancestral great ape acrocentric chromosomes (chimpanzee chromosomes 12 and 13) (Ijdo et al., 1991a; Martin et al., 2002; Fan et al., 2002). The human 2q13 band, near the centromere, contains telomeric repeats in a head-to-head orientation. Over 50 interstitial telomeres have been described (Azzalin et al., 2001).
- Chromosomal regions can become inverted. For example, there are five distinct subtypes of the mosquito *Anopheles gambiae* having varying kinds of paracentric inversions on chromosome 2 (Holt et al., 2002). In humans, an inversion of a single gene causes a severe form of hemophilia (Antonarakis et al., 1995).
- A portion of a chromosome may be deleted. Deletions may be terminal or interstitial; an example of a terminal deletion of chromosome 11q is shown in Figure 16.19 (arrow B).
- Segmental duplications commonly occur, as discussed above.
- Normal chromosomes from any eukaryotic species can vary between individuals in length, number, and position of heterochromatic segments.
- Fragile sites often occur, sometimes causing chromosomal breaks. These fragile sites can be inherited in a dominant Mendelian fashion.
- At least some eukaryotes display chromatin diminution, a form of developmentally programmed DNA rearrangement. Remarkably, chromosomes in somatic cells can fragment, then lose some chromosomal material. Thus somatic chromosomes can have a different structural organization and a smaller gene number than germline cells. Chromatin diminution could represent an unusual gene-silencing mechanism (Müller and Tobler, 2000). This phenomenon has been observed in at least 10 nematode species, including the horse intestinal parasite *Parascaris univalens* (also called *Ascaris megalocephala*) and the hog parasite *Ascaris suum*.

You can read about this hemophilia at the Online Mendelian Inheritance in Man (OMIM) site at NCBI (entry 306700). We will describe OMIM in Chapter 18.

Deletion 11q syndrome results in trigonencephaly (a triangle-shaped head), a carp-shaped mouth, and cardiac defects (Jones, 1997).

Comparison of Eukaryotic DNA: PipMaker and VISTA

Comparative genomics is a powerful approach to annotating and interpreting the meaning of genomic DNA from multiple organisms. When we analyze the genomes of organisms that diverged recently (e.g., humans and chimpanzees diverged 5 MYA) or in the distant past (e.g., mosquitoes and fruit flies diverged 250 MYA; Zdobnov et al., 2002), it is helpful to align the genomic sequences in order to define conserved regions. Such analyses can provide a wealth of information about the existence and evolution of protein-coding genes and other DNA features as well as information about chromosomal evolution.

Genes from different organisms that are derived from a common ancestor and that share a common function are called orthologs (Chapter 3). In comparing genomic sequences from two (or more) organisms, we may wish to analyze regions in each species having orthologous genes. Such regions are said to have conserved synteny. Synteny denotes the occurrence of two or more gene loci on the same chromosome, regardless of whether or not they are genetically linked. This definition refers to an arrangement of genes along a chromosome within a single species. "Conserved synteny" refers to the occurrence of orthologous genes (i.e., in two species) that are syntenic. As an example, the occurrence of the neighboring genes *RBP4* and *CYP26A1* on human chromosome 10 and mouse chromosome 19 represents conserved synteny (see Fig. 16.27 below).

Synteny derives from Greek roots meaning "same thread" or "same ribbon." A common error is to refer to orthologous genes as being syntenic when instead they share conserved synteny (Passarge et al., 1999).

In order to analyze regions of conserved synteny—or even larger regions of genomic DNA that do not necessarily contain protein-coding genes—it is necessary to perform pairwise alignment and multiple sequence alignment of genomic DNA. We discussed approaches to this for prokaryotes (Chapter 14), and we discussed algorithms such as BLAT that are useful for the comparison of large DNA queries to databases containing genomic DNA (Chapter 5).

Two principal tools for the comparison of genomic DNA in eukaryotes are PipMaker (Schwartz et al., 2000) and VISTA (Mayor et al., 2000; reviewed in Frazer et al., 2003). The goal of each program is to align long sequences (e.g., thousands to millions of base pairs) while visualizing conserved segments (exons and presumed regulatory regions) as well as large-scale genomic changes (inversions, rearrangements, duplications). It is important to learn both the order and orientation of conserved sequence features.

The National Institutes of Health Intramural Sequencing Center (NISC) is currently sequencing BAC contig genomic DNA from dozens of genomic regions of conserved synteny in a variety of species, including human, baboon, chimpanzee, cow, mouse, rat, dog, cat, chicken, zebrafish, and *Fugu*. See ▸http://www.nisc.nih.gov.

PipMaker and MultiPipMaker are available at ▸http://bio.cse.psu.edu/pipmaker/. ("Pip" stands for "percent identity plot.") VISTA (Visualization Tools for Alignments) is at ▸http://www-gsd.lbl.gov/vista/. mVISTA (main VISTA) is a program for visualizing genomic alignments, while rVISTA (regulatory VISTA) is used to align transcription factor binding sites. AVID is an alignment algorithm used by the VISTA tools (Bray et al., 2003). The Berekeley Genome Pipeline includes a VISTA browser (▸http://pipeline.lbl.gov/). This allows human–mouse, mouse–rat, and human–rat genomic DNA comparisons.

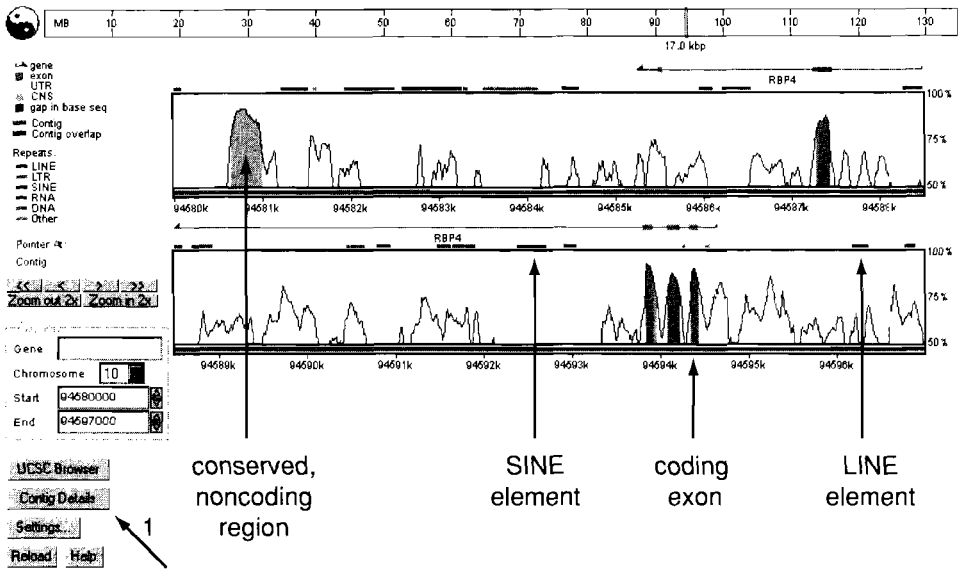
As an example of how to perform genomic DNA alignments, go to the VISTA browser at Berkeley and enter *RBP4*. The result shows the sequence (in bases) on human chromosome 10 (x axis) and the percent nucleotide identity between human and mouse (y axis) (Fig. 16.20). The color-coded output identifies features such as putative coding exons, conserved noncoding regions, and repetitive DNA elements. PipMaker and MultiPipMaker provide similar outputs. An example is shown in Fig. 16.21, where 100,000 nucleotides of human chromosome 10 and the corresponding region of mouse chromosome 19 were used as input (in the FASTA format). The output includes a dot plot (not shown), an overview of the aligned regions (Fig. 16.21*a*), a pairwise alignment (Fig. 16.21*b*), and summary statistics (Fig. 16.21*c*). Programs such as VISTA and PipMaker will have an increasingly essential role in genome analysis.

INDIVIDUAL EUKARYOTIC GENOMES

Introduction

We turn next to a survey of eukaryotic genomes. For our analysis of all genomes, the main issues include (1) description of the complete sequence of each chromosome, (2) annotation of the DNA to identify and characterize noncoding DNA, and (3) identification of protein-coding genes and other noncoding genes. Further issues include (1) analyses of chromosome structure, such as regions of duplication or deletion; (2) comparative analyses with the genomes of related organisms as part of an attempt to reconstruct the molecular evolutionary history of species; and

(*a*)



(*b*)

```
Criteria: 70% identity over 100 bp

*************** Conserved Regions  ***************

94585364        to    94585486    =    129bp      at    69.80% UTR
94594441        to    94594458    =    18bp       at    72.20% UTR
94594583        to    94594652    =    70bp       at    81.40% UTR
94587237        to    94587445    =    209bp      at    83.30% exon
94593805        to    94593910    =    106bp      at    91.50% exon
94594080        to    94594215    =    136bp      at    86.80% exon
94594331        to    94594440    =    110bp      at    90.90% exon
94589637        to    94589864    =    229bp      at    72.50% noncoding
94589940        to    94590050    =    112bp      at    69.60% noncoding
94590435        to    94590544    =    111bp      at    73.00% noncoding
94591250        to    94591381    =    133bp      at    73.70% noncoding
94593365        to    94593457    =    93bp       at    72.00% noncoding
```

FIGURE 16.20. *The VISTA program for aligning genomic DNA sequences is available through a web browser that can be queried with text or DNA sequence (up to 300,000 bases).* (a) *The output for an* RBP4 *text query is shown here. The* x *axis shows the nucleotide position along human chromosome 10, and the* y *axis shows the percent nucleotide identity between human and mouse. A variety of coding and noncoding features are indicated. There is a link to the UCSC Genome Browser (described in Chapter 17). By clicking on Contig Details (arrow 1),* (b) *a summary of aligned segments is provided.*

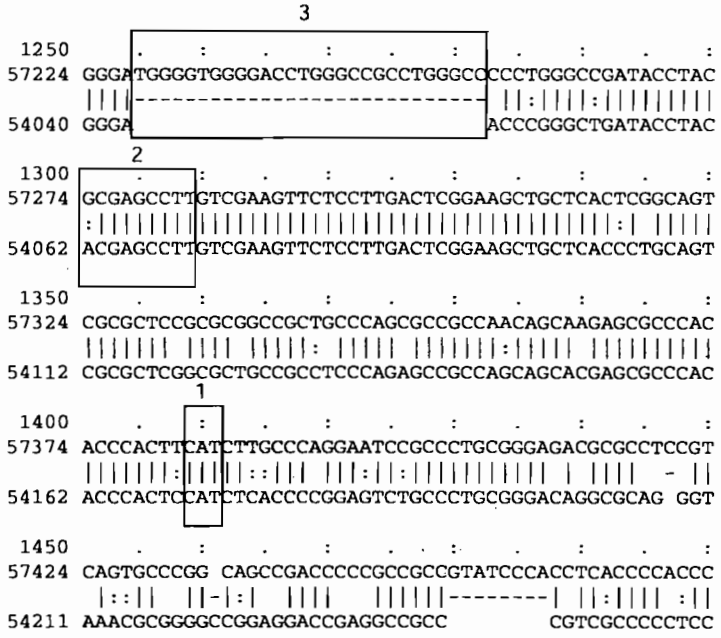
(3) strategies to relate the genomic DNA sequence to the phenotype of the organism. This phenotype includes an organism's strategies for adaptation to its environment, evolution, metabolism, growth, development, maintenance of homeostasis, and reproduction.

In the case of many eukaryotes—from the protozoans such as *Plasmodium* to pathogenic fungi and parasitic worms—we also want to understand the genetic basis of how the organism causes disease and how we can counterattack. At present, there are no vaccines available to prevent diseases caused by any eukaryotic parasites that infect humans, including protozoans (such as trypanosomes) and helminths (parasitic nematodes). The availability of whole genome sequences may provide clues as to which antigens are promising targets for vaccine development and

The word protozoan derives from the Greek *proto* ("early") and *zoion* ("animal"). This contrasts with the word metazoan (animal) from the Greek *meta* ("after"; at a later stage of development) and *zoion*.

(a)
```
Aligned to ">gi|20889106:4335093-4435093 Mus musculus WGS supercontig
Mm19_WIFeb01_321":
:
1040-1099 <--> 36101-36160 73% (60 nt)
1100-1128 <--> 36168-36196 55% (29 nt)
:
1228-1239 <--> 9578-9589 92% (12 nt)
1240-1255 <--> 9598-9613 44% (16 nt)
1257-1296 <--> 9614-9653 68% (40 nt)
1299-1301 <--> 9654-9656 100% (3 nt)
1311-1328 <--> 9657-9674 61% (18 nt)
1329-1505 <--> 9676-9852 62% (177 nt)
1508-1524 <--> 9853-9869 76% (17 nt)
:
```

(b)



(c)
```
Local Alignment Number 317
Similarity Score:  92303
Match Percentage:  62 %
Number of Matches:  1985
Number of Mismatches:  752
Total Length of Gaps:  456
Begins at (55993,52948) and Ends at (59050,55819)
```

FIGURE 16.21. *The PipMaker program performs pairwise sequence alignment of large stretches of genomic DNA: a portion of human chromosome 10 and mouse chromosome 19 (gi 20889106). The alignment includes* (a) *a tabular summary, the alignment [a portion is shown in* (b)*], and* (c) *a summary. These regions contain the* RBP4 *gene on the bottom strand orientation. The start codon (ATG) appears on the bottom strand as 5′ TAC 3′ (box 1). The final nine nucleotides of human RBP4 exon 2 (5′ AAGGCTCGC 3′ encoding the amino acids lysine-alanine-arginine) can be seen as the reverse complement (5′ GCGAGCCTT 3′) in the human sequence (box 2). This is followed by a downstream gap in the mouse genomic sequence in an intronic region (box 3). The human sequence in this figure can be compared to Figure 6.4.*

pharmacological intervention. For example, predicted secreted surface proteins can be expressed in bacteria and used to immunize mice in order to develop potential vaccines (Fraser et al., 2000).

As we begin our description of eukaryotic genomes, we will refer to the phylogenetic tree shown in Figure 16.1. We will describe genome-sequencing projects from the bottom of this tree upward.

A phylogenetic description of the eukaryotes is essential for our understanding of both evolutionary processes that shaped the development of species and the diversity of life today. Evolutionary reconstructions that are based on molecular sequence data typically use small-subunit ribosomal RNA because it has many sites that are phylogenetically informative across all life forms (Van de Peer et al., 2000). We saw an example of such a tree in Figure 12.2. However, there is no uniform consensus on the optimal approach to making a tree (Box 16.1 and Chapter 11).

BOX 16-1
Inconsistent Phylogenies

It is important to note that many phylogenetic reconstructions are inconsistent with each other. There are three main sources of conflicting results (Philippe and Laurent, 1998):

1. Gene duplication followed by random gene loss can cause artifacts in tree reconstruction. This occurred at the whole-genome level in yeast (Chapter 15) and other eukaryotes such as plants and fish (see below).
2. Lateral gene transfer can confuse phylogenetic interpretation (Chapter 14).
3. The technical artifact of long branch chain attraction can confuse phylogenetic analyses. This is a phenomenon where the longest branches of a tree are grouped together, regardless of the true tree topology (Fig. 11.23). It is essential to account for differences in substitution rates among sites within a molecule. Reyes et al. (2000) consider this problem in their phylogeny of the order Rodentia.

Researchers often overcome these potential problems by concatenating multiple protein (or nucleic acid) sequences. For example, the tree in Figure 16.1 is based on four concatenated proteins. Wang et al. (1999) used 75 genes in their comprehensive phylogeny of animals, plants, and fungi (described below); Kumar and Hedges (1998) studied 658 genes in 207 vertebrate species. In another strategy, several groups have made trees based on gene content or gene fusion events (Snel et al., 1999; Stechmann and Cavalier-Smith, 2002).

The diplomonadida are also called diplomonads. This group includes the family Hexamitidae, which further includes the genus *Giardia* (described below).

The microsporidia such as *Encephalitozoon* used to be classified as deep-branching eukaryotes. Subsequent analysis of the complete *E. cuniculi* genome revealed that this microsporidial parasite is closely related to the fungi (Chapter 15 and Fig. 16.1).

Giardia was the first parasitic protozoan of humans observed with a microscope by Antony van Leeuwenhoek (in 1681).

The U.S. Food and Drug Administration offers information on *Giardia* at ▸http://vm.cfsan.fda.gov/~mow/chap22.html.

Organisms that lack peroxisomes could provide us insight into fatty acid metabolism or other metabolic processes. This in turn could prove helpful to our understanding of human diseases that affect such organelles. The most common human genetic disorder affecting peroxisomes is adrenoleukodystrophy, caused by mutations in the *ABCD1* gene (RefSeq accession NM_000033). Does *Giardia* have an ortholog of this gene?

Protozoans at Base of Tree

The eukaryotes include deep-branching protozoan species from the diplomonadida (such as *Giardia* and *Trichomonas*), discicristata (e.g., *Euglena, Leishmania,* and *Trypanosoma*), alveolata (e.g., *Toxoplasma* and *Plasmodium*), and heterokonta (Fig. 16.1).

There is strong evidence that mitochondrial genes, present in most eukaryotes, are derived from an α-proteobacterium (see Chapter 14). Previously, it was hypothesized that deep-branching organisms such as *Giardia* and *Trichomonas* lack mitochondria. They were thought to have evolved from other eukaryotes prior to the symbiotic invasion of an α-proteobacterium. However, recent analyses of *Giardia, Trichomonas,* and microsporidia such as *Trachipleistophora hominis* suggest the presence of mitochondrial genes (Embley and Hirt, 1998; Williams et al., 2002; Lloyd and Harris, 2002).

Giardia lamblia (also called *Giardia intestinalis*) is a protozoan, water-borne parasite that lives in the intestines of mammals and birds (Adam, 2001). It is the cause of giardiasis, the most frequent source of nonbacterial diarrhea in North America. Like some other unicellular protozoans, *Giardia* lack not only mitochondria but also peroxisomes (responsible for fatty acid oxidation) and nucleoli. Thus the genome of *Giardia* could reflect the adaptations that led to the early emergence of eukaryotic cells.

The *Giardia* genome is approximately 12 Mb (Adam, 2001) and is currently being sequenced by the whole-genome shotgun method (McArthur et al., 2000). Each cell has two morphologically identical nuclei, each nucleus having five chromosomes ranging from 0.7 to over 3 Mb. While the genome sequencing is still in progress, an online database describes currently annotated ORFs.

As we consider the genomes of various eukaryotes, a consistent theme is that transposable elements are extremely abundant, occupying half the entire human

genome (Chapter 17) and causing massive genomic rearrangements. Thus, in order to understand their origins and their function, it is of interest to find eukaryotes that lack these elements. *Giardia* provides such an example. Arkhipova and Meselson (2000) examined 24 eukaryotic species for the presence of two major classes of transposable elements (retrotransposon reverse transcriptases and DNA transposons). They found them present in all species except bdelloid rotifers, an asexual animal (see Table 16.22 below). Deleterious transposable elements thrive in sexual species, but they are unlikely to propagate in asexual species because of strong selective pressure against having active elements. Further inspection of the asexual *Giardia* by Arkhipova and Morrison (2001) revealed just three retrotransposon families. One of these is inactive, and the other two are telomeric. This location could provide a buffer between protein-coding genes and the telomeres, and these elements could contribute to the ability of *Giardia* to vary the length of its chromosomes in response to environmental pressures—for example, chromosome 1 can expand from 1.1 to 1.9 Mb (Pardue et al., 2001).

The *Giardia* genome project website is at ▸http://www.mbl.edu/Giardia (McArthur et al., 2000).

Another basic question about eukaryotic genomes is the origin of introns. Spliceosomal introns occur commonly in the "crown group" of eukaryotes (the kingdoms Animalia, Plantae, and Fungi). However, their presence in the earliest branching protozoa has been disputed (Johnson, 2002), and introns have not been detected in parabasalids such as *Trichomonas*. Nixon et al. (2002) identified a 35-bp intron in a gene encoding a putative [2Fe-2S] ferredoxin. Simpson et al. (2002) subsequently identified several introns in *Carpediemonas membranifera,* a eukaryote thought to be a close relative of *Giardia*. These findings suggest that if introns were a eukaryotic adaptation, they arrived early in evolution and possibly in the last common eukaryotic ancestor.

You can study this *Giardia* ferredoxin gene at GenBank (DNA accession AF393829). To find the intron, try using BLAST 2 Sequences (Chapter 3) to compare the protein (or the DNA encoding the protein) to the genomic DNA.

Genomes of Unicellular Pathogens: Trypanosomes and *Leishmania*

There are about 20 species in the protozoan genus *Trypanosoma* (reviewed in Donelson, 1996). Two of these are pathogenic in humans (Cox, 2002). *Trypanosoma brucei* subspecies cause several forms of sleeping sickness, a fatal disease that infects hundreds of thousands of people in Africa. *Trypanosoma cruzi* causes Chagas' disease, prevalent in South and Central America. The adverse impact of these trypanosomes is even greater because they also afflict livestock. Tsetse flies or other insects transmit the trypanosomes to humans.

Tsetse flies are insects that feed on vertebrate blood. To obtain additional nutrients beyond what is available in blood, tsetse flies harbor two obligate intracellular bacteria: *Wigglesworthia glossinidia* and *Sodalis glossinidius*. The *W. glossinidia* genome was recently sequenced (see RefSeq accession NC_004344). Similar to other intracellular bacteria (Chapter 14), it has a reduced genome size of only 700,000 bp (Akman et al., 2002).

The African trypanosome genome network (Table 16.12) is coordinating the sequencing of trypansome genomes. The genome of *T. brucei* is 35 Mb, although its size varies by up to 25% in different isolates (reviewed in El-Sayed et al., 2000). There are at least 11 pairs of large, diploid, nuclear chromosomes (ranging in size form about 1 Mb to >6 Mb). Additionally, there are variable numbers of intermediate chromosomes (200–900 kb), and there are about 100 linear minichromsomal DNA

TABLE 16-12 Web Resources for *Trypanosome* Genomics

Resource	Comment	URL
The *Trypanosoma brucei* Genome Network	Sponsored by the Wellcome Trust	▸http://parsun1.path.cam.ac.uk/
Trypanosoma brucei omni Blast Server	Sanger Institute	▸http://www.sanger.ac.uk/Projects/T_brucei/Toolkit/blast_server.shtml
Trypanosoma cruzi Genome Initiative Information Server	From the Oswaldu Cruz Institute, Brazil	▸http://www.dbbm.fiocruz.br/TcruziDB/index.html

TABLE 16-13 Web Resources for *Leishmania* Genomics

Resource	Comment	URL
The *Leishmania major* Friedlin Genome Project	At the Wellcome Trust Sanger Institute	▸http://www.sanger.ac.uk/Projects/L_major/
Seattle Biomedical Research Institute (SBRI)	Information on various infectious disease	▸http://www.genome.sbri.org/
The European *Leishmania major* Friedlin Genome Sequencing Consortium	A listing of participating laboratories	▸http://www.sanger.ac.uk/Projects/L_major/EUseqlabs.shtml

molecules (50–150 kb). Some of these minichromosomes contain a 177-bp repeat that comprises more 90% of the total sequence (El-Sayed et al., 2000).

Another remarkable feature of trypanosomes is the presence of a massive network of circular rings of mitochondrial DNA, termed kinetoplast DNA. Thousands of rings of kinetoplast DNA interlock in a shape resembling medieval armor (Shapiro and Englund, 1995). Kinetoplast DNA occurs as maxicircles (present in several dozen copies) and minicircles (present in thousands of copies). These include a universal minicircle sequence of 12 nucleotides that serves as a replication origin (Morris et al., 2001).

For information on sleeping sickness, see the World Health Organization website at ▸http://www.who.int/tdr/diseases/tryp/default.htm.

See problem 16.3 for an exercise on a trypanosome universal minicircle binding protein. For an example of a maxicircle sequence and the genes it encodes, see GenBank accession M94286.

Leishmania major is another deadly protozoan parasite in the Euglenozoa (Fig. 16.1). Thirteen different species of *Leishmania* cause the disease leishmaniasis, for which there is no effective vaccine and limited pharmacological intervention available. These various *Leishmania* species have from 34 to 36 chromosomes (Myler et al., 2000).

The *Leishmania major* genome is about 34 Mb with 36 chromosomes (from 0.3 to 2.5 Mb). The *Leishmania* Genome Network is coordinating the sequencing of the genome (Table 16.13). The nucleotide sequence was determined for chromosome 1 (the smallest chromosome) and was found to have a remarkable genomic organization (Myler et al., 1999). The first 29 genes (from the left telomere) are all transcribed from the same DNA strand, while the remaining 50 genes are all transcribed from the opposite strand. This polarity is unprecedented in eukaryotes and resembles prokaryote-like operons. It has a 257-kb region that is filled with 79 protein-coding genes (≈1 gene each 3200 bp). Myler et al. (1999) estimate that overall there are about 9800 genes encoded in the genome.

The World Health Organization offers information on leishmaniasis at ▸http://www.who.int/inf-fs/en/fact116.html and ▸http://www.who.int/tdr/diseases/leish/default.htm.

Malaria Parasite *Plasmodium falciparum* and Other Apicomplexans

Malaria kills about 2.7 million people each year, mostly children in Africa, and almost 500 million people are newly infected each year. It is caused by the apicomplexan parasite *Plasmodium falciparum*. While there are 120 species of *Plasmodium*, only four infect humans: *P. falciparum, P. vivax, P. ovale,* and *P. malariae*. The main vector for malaria in Africa is the mosquito, *A. gambiae*.

The alveolates are protists that include three branches: (1) the phylum Apicomplexa includes parasites such as *P. falciparum* and *Toxoplasma gondii,* (2) the ciliates include *Paramecium* and *Tetrahymena thermophila,* and (3) the dinoflagellates include a cause of paralytic shellfish poisoning, *Alexandrium*.

The name Apicomplexa derives from a characteristic apical complex of microtubules. You can read more about apicomplexans online at ▸http://www.ucmp.berkeley.edu/protista/apicomplexa.html or ▸http://www.tulane.edu/~wiser/protozoology/notes/api.html.

Plasmodium falciparum has a complex lifestyle, contributing to the challenge of developing a successful vaccine (Cowman and Crabb, 2002; Wirth, 2002; Long and Hoffman, 2002). *Plasmodium* resides in the salivary glands and gut of the mosquito *A. gambiae*. When a mosquito bites a human, it introduces the parasite in the sporozoite form that infects the liver. *Plasmodium* then matures to the merozoite form, which attaches to and invades human erythrocytes through host cell receptors. Within erythrocytes, trophozoites form. Some merozoites transform into gametocytes, which are captured when mosquitoes feed on infected individuals. A goal of sequencing the *P. falciparum* genome is to find gene products that function at

TABLE 16-14 Genomics Resources for *Plasmodium falciparum* and Malaria

Resource	Comment	URL
PlasmoDB	Main web resource for *P. falciparum*	▸http://www.plasmodb.org/
Links page	At NCBI	▸http://www.ncbi.nlm.nih.gov/projects/Malaria/related_links.html
P. falciparum Genome Project	At the Sanger Institute	▸http://www.sanger.ac.uk/Projects/P_falciparum/

selective stages of the parasite life cycle, offering targets for drug therapy or vaccine development.

The complete genome sequence of *P. falciparum* was reported by an international consortium (Gardner et al., 2002). The sequencing was extraordinarily challenging because the AT (adenine and thymine) content of the genome is 80.6% overall, which is the highest for any eukaryotic genome. In intergenic regions and introns, the AT content reached 90% in some cases. A whole-chromosome (rather than a whole-genome) shotgun sequencing strategy was employed. With this approach, chromosomes were separated on pulsed-field gels, DNA was extracted, and shotgun libraries containing 1–3 kb of DNA were constructed and sequenced. The genome is 22.8 Mb, with 14 chromosomes from 0.6 to 3.3 Mb.

Gardner et al. (2002) identified 5268 protein-coding genes in *P. falciparum*. This is the same number as is predicted for *Schizosaccharomyces pombe* (Chapter 15), although the genome size is twice as large. There is one gene approximately every 4300 bp overall. Gene Ontology Consortium terms (Chapter 8) were assigned to about 40% of the gene products(≈2100). However, about 60% of the predicted proteins have no detectable homology to proteins in other eukaryotes. These proteins are potential targets for drug therapies. For example, some are essential for the function of the apicoplast. This is a relic plastid, unique to Apicomplexa and homologous to the chloroplast, that functions in fatty acid and isoprenoid biosynthesis.

There are several main resources on the web to study *P. falciparum* (Table 16.14). PlasmoDB (Bahl et al., 2003) is the centralized resource for genomic data.

In parallel to the *P. falciparum* genome project, a consortium sequenced the genome of the rodent malaria parasite, *Plasmodium yoelii yoelii* (Carlton et al., 2002). This is an extremely important accomplishment because the complete life cycle of *P. falciparum* cannot be maintained in vitro, while *P. yoelii yoelii* can. This genome is 23.1 Mb and has 14 chromosomes, as does *P. falciparum*. The AT content is comparably high (77.4%). The genomes are also predicted to encode a comparable number of genes. When the full set of predicted *P. falciparum* proteins (5268) were searched against the predicted *P. yoelii yoelii* proteins (5878 proteins) by BLAST searching (with an E value cutoff of 10^{-15}), 3310 orthologs were identified. These include vaccine antigen candidates known to elicit immune responses in exposed humans (Carlton et al., 2002).

Having the genome sequences of *P. falciparum* and *P. yoelii yoelii* available, how can bioinformatics and genomics approaches be used to understand the basic biology of these organisms? Data are now available on thousands of previously unknown genes, offering many new potential strategies to combat malaria (Hoffman et al., 2002).

- The apicoplast is a potential drug target. Zuegge et al. (2001) analyzed the amino terminal sequences of 84 proteins targeted to apicoplasts and 102 non-apicoplast (e.g., cytoplasmic, secretory, or mitochondrial) sequences. They

For online facts on malaria, see ▸http://www.wellcome.ac.uk/en/malaria/ and ▸http://www.who.int/tdr/diseases/malaria/default.htm.

Charles Louis Alphonse Laveran won a Nobel Prize in 1907 for his work on malaria-causing trypanosomes (▸http://www.nobel.se/medicine/laureates/1907/index.html). Earlier, Ronald Ross was awarded a Nobel Prize for his studies of malaria (▸http://www.nobel.se/medicine/laureates/1902/index.html).

The *P. falciparum* genome was sequenced by a consortium including the Wellcome Trust Sanger Institute, The Institute for Genomic Research, The U.S. Naval Medical Research Center (NMRC, Maryland), and Stanford University. The genome of the slime mold *Dictyostelium discoideum* also has a high AT content (see below).

For an NCBI website on malaria genetics and genomics, visit ▸http://www.ncbi.nih.gov/projects/Malaria/.

A plastid is any photosynthetic organelle. The most well known plastid is the chloroplast, found in green algae and land plants (Gilson and McFadden, 2001). See the section on plants below.

Isoprenes are five-carbon chemical molecules that combine to form many thousands of natural compounds, including steroids, retinol, and odorants. (RBP and OBP are lipocalins that transport isoprenoids.)

used principal components analysis, neural networks, and self-organizing maps (Chapter 7) to build a predictive model for apicoplast targeting signals.

- Comparative genomics approaches yield important insight into the genome structure, gene content, and other genomic features of closely related species. Carlton et al. (2001) compared ESTs and genome survey sequences (see Chapter 2) from *P. falciparum, P. vivax,* and *P. berghei.* As part of this analysis, they identified the most highly expressed genes, such as the *rif* gene family of *P. falciparum* that is implicated in antigenic variation.
- A map of conserved syntenic regions between *P. yoelii yoelii* and *P. falciparum,* covering over 16 Mb overall, provides insight into the evolution of these parasites. Carlton et al. (2002) used the MUMmer program (Chapter 14) to align protein-coding regions. The conserved synteny map reveals regions of conserved gene order, allows analysis of chromosomal break points, and confirms the absence of some genes (such as *var* and *rif* in *P. yoelii yoelii*).
- Genes that function in antigenic variation and immune system evasion can be investigated. In *P. vivax,* there are as many as 1000 copies of *vir,* a gene family localized to subtelomeric regions. *Plasmodium yoelii yoelii* has 838 copies of a related gene, *yir* (Carlton et al., 2002).
- Several groups applied proteomics approaches to analyze the proteins of *P. falciparum* at four stages of the life cycle (sporozoites, merozoites, trophozoites, and gametocytes). Florens et al. (2002) identified 2415 expressed proteins, about half of which are annotated as hypothetical. An unexpected finding was that the *var* and *rif* genes—thought to be involved in immune system invasion—were abundantly present in the sporozoite stage. Together, these studies define stage-specific expression of proteins, suggesting possible protein functions. Proteomics approaches also validate the gene-finding approaches from genomic DNA. Lasonder et al. (2002) identified some protein sequences by mass spectrometry that were not initially predicted using gene-finding algorithms to analyze genomic DNA.
- It is possible to identify *Plasmodium* metabolic pathways as therapeutic targets (Gardner et al., 2002; Hoffman et al., 2002). All organisms studied to date synthesize isoprenoids using isopentyl diphosphate as a building block. An atypical pathway employed by some plants and bacteria involves 1-deoxy-D-xylulose 5-phosphate (DOXP). This DOXP pathway is absent in mammals. Jomaa et al. (1999) used TBLASTN (with a bacterial DOXP reductoisomerase protein as a query against a *Plasmodium* genomic DNA database) and found an orthologous *Plasmodium* gene. They showed that this protein is likely localized to the apicoplast and that *P. falciparum* survival is sensitive to low levels of two inhibitors of the enzyme. They further showed that these drugs have antimalarial activity in mice infected with *Plasmodium vinckei.* This type of bioinformatics-based approach holds great promise in the search for additional antimalarial drugs.

We have briefly described several of the protozoan genome sequence projects. Additional projects are in progress for additional protozoans such as the ciliate *Paramecium* (Dessen et al., 2001) and alveolates of the genus *Tetrahymena.* The Institute for Genomic Research, which is involved in sequencing many prokaryotic genomes (Chapter 14), is also sequencing a variety of protozoan and other eukaryotic genomes.

Genome-sequencing projects are underway for additional species that cause malaria in rodents, including *Plasmodium berghei* and *Plasmodium chabaudi* (Hoffman et al., 2002).

The Prediction of Apicoplast Targeted Sequences (PATS) database is available at ▸http://gecco.org.chemie.uni-frankfurt.de/pats/pats-index.php.

We encountered *vir* in Chapter 5 (problem 5.2) where we used both BLAST and PSI-BLAST to evaluate the family.

The *Paramecium* genome project website is at ▸http://paramecium.cgm.cnrs-gif.fr/ while that for *Tetrahymena* is at ▸http://lifesci.ucsb.edu/~genome/Tetrahymena/. You can see the phylogenetic placement of these organisms in Figure 16.1. The *T. gondii* database ToxoDB is available at ▸http://ToxoDB.org (Kissinger et al., 2003).

Eukaryotic genome projects from TIGR (▸http://www.tigr.org/tdb/euk/) include a variety of protozoan (and metazoan) parasites such as *P. falciparum, P. vivax, P. yoelii, T. gondii, T. brucei, T. cruzi,* and *Schistosoma mansoni.*

FIGURE 16.22. *The evolution of plants, animals, and fungi. The estimated time of divergence of plants, fungi, and animals is 1.5 BYA according to a phylogenetic study (adapted from Wang et al., 1999). Prior to this divergence event, a single-celled eukaryotic organism acquired an α-proteobacterium (the modern mitochondrion, present today in animals, fungi, and plants). After the divergence of plants from animals and fungi about 1.5 BYA, the plant lineage acquired a plastid (the chloroplast). According to this model, metazoans diverged about 400 million years earlier than predicted by the fossil record. Also, nematodes (e.g.,* C. elegans*) diverged earlier than chordates (e.g., vertebrates) and arthropods (e.g., insects). Adapted from separate studies by Meyerowitz (2002) and Wang et al. (1999). Used with permission.*

Plant Genomes

Overview

Hundreds of thousands of plant species occupy the planet. Molecular phylogeny shows us that plants form a distinct clade within the eukaryotes (see Viridiplantae, Fig. 16.1). All plants are multicellular because they develop from embryos, which are multicellular structures enclosed in maternal tissue (Margulis and Schwartz, 1998). Most plants have the capacity to perform photosynthesis, although some (such as the beech drop, *Epifagus*) do not.

The analysis of plant genomes allows us to address the molecular genetic basis of characteristics that distinguish plants from animals such as the presence of specialized cell walls, vacuoles, plastids, and cytoskeleton. Plants are sessile and depend on photosynthesis. The sequencing of plant genomes is likely to lead to explanations for many of these basic features.

When did the lineages leading to today's plants diverge from animals, fungi, and other organisms? The earliest evidence of life is from about 3.8 billion years ago (BYA), while eukaryotic fossils have been dated to 2.7 BYA. These events are depicted in the schematic tree of Figure 16.22, based on separate studies by Meyerowitz (2002) and Wang et al. (1999). There are no very early plant fossils extant, and it is thus difficult to assess the dates that species diverged from each other. Various researchers have used molecular clocks based on protein, DNA (nuclear or mitochondrial), or RNA data. A study by Wang et al. (1999) used a combined analysis of 75 nuclear genes to estimate the divergence times of plants, fungi, and several animal phyla. Their estimates of divergence time were calibrated based on evidence from the fossil record that birds and mammals diverged 310 MYA. They found that animals and plants diverged 1547 MYA, at almost exactly the same time that animals and fungi diverged (1538 MYA) (Fig. 16.22).

The *Epifagus virginiana* chloroplast genome has been sequenced (NC_001568) (Wolfe et al., 1992). *Epifagus* is parasitic on the roots of beech trees. The original major function of its chloroplast genome, photosynthesis, has become obsolete. It lacks six ribosomal protein and 13 tRNA genes that are present in the chloroplast genomes of photosynthetic plants (Wolfe et al., 1992).

Plants and animals differ greatly in their gene content. For example, plants lack intermediate filaments and the genes that encode intermediate filament *proteins* such as cytokeratin and vimentin.

The use of 18S RNA has suggested an animal–fungi clade (Fig. 16.1), consistent with Figure 16.22.

BOX 16-2
An Astonishing Way to Acquire a Chloroplast: Eat an Organism That Has It

The chloroplast is a plastid (photosynthetic organelle) in plants that contains the green pigment chlorophyll. Chloroplasts convert light to energy. A major hypothesis about their origin is that a eukaryotic cell acquired a cyanobacterium soon after the divergence of plants from animals and fungi (Fig. 16.22). But a radically different mechanism is also common: A eukaryote can ingest an alga (i.e., another eukaryote) that already has a chloroplast (Gilson and McFadden, 2002). This process, called endosymbiosis, may have occurred independently in at least seven separate eukaryotic groups: apicomplexa (discussed above), chlorarachniophytes, cryptomonds, dinoflagellates, euglenophytes, heterokonts, and haptophytes (reviewed in Gilson and McFadden, 2002).

Most chloroplast-containing plants and algae have three genomes in each cell: a nuclear genome, a mitochondrial genome, and a chloroplast genome. In cryptomonads (such as *Guillardia theta*) and chlorarachniophytes, there is an additional, fourth distinct genome: the vestigial nuclear genome of the engulfed alga. This second nucleus is called a nucleomorph.

Just as the genome of intracellular bacteria is highly reduced, the nucleomorph genome is extremely small. Douglas et al. (2001) sequenced the nucleomorph genome of *G. theta*. It is only 551,264 bp. The gene density is extraordinarily high, with one gene per 977 bp. The noncoding regions are extremely short, and there is only one pseudogene. Some genes, such as those encoding DNA polymerases, are absent and the gene product must be imported to the plastid across four separate membranes.

The circlular plastid DNA of *G. theta* is also very compacted. Douglas and Penny (1999) sequenced this genome of 121,524 bp and found that 90% of the DNA is coding, with no pseudogenes or introns. (In contrast, only 68% of the rice plastid genome is coding.) You can explore the *G. theta* plastid genome at NCBI (accession NC_000926) and compare it with the plastid genome of the red alga *Pophyra purea*, a rhodophyte (accession NC_000925). These two genomes show a high degree of conserved synteny. You can also compare the *G. theta* plastid genome to that of the diatom *Odontella sinensis* (accession NC_001713). This is a related alga that also acquired its plastid by secondary endosymbiosis but lacks a nucleomorph.

The earliest known plant fossils date from the Silurian period (430–408 MYA) (Margulis and Schwartz, 1998).

Angiosperms are flowering plants in which the seeds are enclosed in an ovary that ripens into a fruit. Monocots are characterized by an embryo with a single cotyledon (seed leaf); examples are rice, wheat, and oats. Eudicots (also called dicotyledons) have an embryo with two seed leafs; examples are tomato and potato. Eudicots include the majority of flowers and trees (but not conifers). The Angiosperm Phylogeny website is at ▸http://www.mobot.org/MOBOT/Research/APweb/welcome.html. It includes dozens of phylogenetic trees, with access to text, photographs of plants, and extensive references. In contrast to angiosperms, gymnosperms develop their seeds in cones.

The early appearance of plants, animals, and fungi may have occurred with the divergence of a unicellular progenitor. Thus a comparison of plants and animals allows us to see how plants and animals independently evolved into multicellular forms (Meyerowitz, 2002). The mitochondrial genes of plants and animals are homologous, indicating that their common ancestor was invaded by an α-proteobacterium (Fig. 16.22). After their divergence, in another endosymbiotic event, a cyanobacterium occupied plant cells to ultimately form the chloroplast. This occurred independently several times (Box 16.2). Still, it has proven difficult to date these events (Meyerowitz, 2002). The first appearance of most animal phyla in the fossil record occurs in many samples dated 530 MYA—the "Cambrian explosion."

We may begin our bioinformatics and genomics approaches to plants by exploring their classification at NCBI (Table 16.15). The relationships of major plant divisions are also illustrated by a phylogenetic tree using sequences of a key plant enzyme, rubisco (Fig. 16.23). The two main groups of Viridiplantae are Chlorophyta (green algae such as the generum *Chlamydomonas*) and Streptophyta. Streptophyta is further subdivided into additional groups such as mosses, liverworts, and the angiosperms (flowering plants), including the familiar monocots and eudicots.

Of the flowering plants, the thale cress *A. thaliana* (a eudicot) and two rice variants (*O. sativa*, a monocot) were the first two genomes to be sequenced. A variety of other plant genome-sequencing projects are in progress (Table 16.16).

Arabidopsis thaliana Genome

Arabidopsis thaliana is a thale cress that is often considered to be a weed. Nonetheless, this organism is prominent as having the first plant genome to be sequenced. *Arabidopsis* has been adopted by the plant research community as a model organism

TABLE 16-15 Classification of Plants (Kingdom Viridiplantae)

The vascular plants are highlighted.

No Rank	Class	Phylum	Order	Example(s)
Chlorophyta (green algae)	Chlorophyceae	–	–	*Chlamydomonas reinhardtii*
	Pedinophyceae			
	Picocystophyceae			
	Prasinophyceae			
	Trebouxiophyceae			
	Ulvophyceae			
Streptophyta	Charophyta/ Embryophyta	Charales	–	*Chara fragilis*
		Coleochaetales	–	*Coleochaete nitellarum*
		Embryophyta (plants)	Anthocerotophyta (hornworts)	*Megaceros aenigmaticus*
			Bryophyta (mosses)	*Physcomitrella patens*
			Marchantiophyta (liverworts)	*Asterella gracilis*
			Tracheophyta (vascular plants)	*Arabidopsis thaliana*
	Chlorokybophyceae			
	Klebsormidiophyceae			
	Mesostigmatophyceae			
	Zygnemophyceae			

Source: Adapted from the NCBI Taxonomy Browser (http://www.ncbi.nlm.nih.gov/)

to study because it is small, has a short generation time, has many offspring, and is convenient for genetic manipulations. It is a member of the Brassicaceae (mustard) family, which includes horseradish, broccoli, cauliflower, and turnips. It is one of about 250,000 species of flowering plants, a group that emerged 200 MYA (Walbot, 2000). Comparative genomics analyses will allow the comparison of the *Arabidopsis* genome to the genomes of other flowering plants in order to learn more about plant genomics (Hall et al., 2002).

The *Arabidopsis* genome is about 125 Mb. Its genome size is thus very small compared to agriculturally important plants such as wheat and barley (16.5 and 5 Gb,

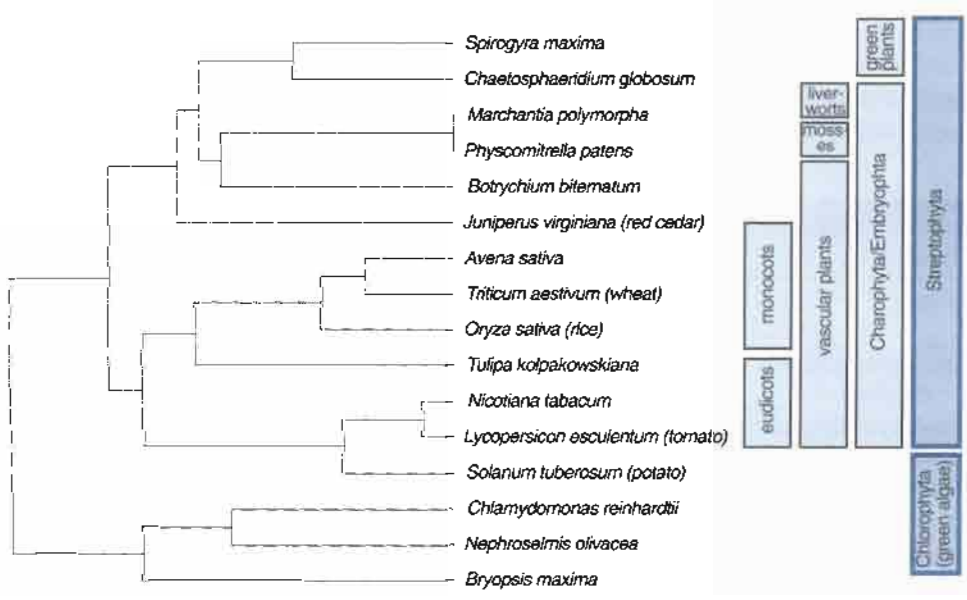


FIGURE 16.23. *A neighbor-joining tree of the plants using rubisco protein. Compare Table 16.15.*

Rubisco is ribulose-1, 5-diphosphate carboxylase. It is an enzyme localized to chloroplasts that catalyzes the first step of carbon fixation in photosynthesizing plants. The enzyme irreversibly converts ribulose diphosphate and carbon dioxide (CO_2) to two 3-phosphoglycerate molecules. The gene name for rubisco is *rbcL*, and for a typical example see the rice protein (RefSeq accession NP.039391).

Online databases are available for model plant genome projects, such as MtDB for *Medicago trunculata* (Lamblin et al., 2003) (http://www.medicago.org/) and ZmDB for maize (Dong et al., 2003) (http://zmdb.iastate.edu/). More comprehensive plant genomics databases include GénoPlante-Info (GPI) (Samson et al., 2003) (http://genoplante-info.infobiogen.fr/) and Sputnik (Rudd et al., 2003) (http://mips.gsf.de/proj/sputnik/). GrainGenes, a database for wheat, barley, rye, and oat, is available at http://www.graingenes.org (Matthews et al., 2003).

TABLE 16-16 Major Plant Genome-Sequencing Projects

See the NCBI plant resources at ▸ http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/PlantList.html. Many plant genomes are about the same size as the human genome (3.2 Gb).

Plant	Common Name	Genome Size (Gb)	Size Relative to Human
Arabidopsis thaliana	Thale cress	0.125	25.6-fold smaller
Avena sativa	Oat	16	5-fold larger
Glycine max	Soybean	1–2	About 2-fold smaller
Hordeum vulgare	Barley	5	1.7-fold larger
Lycopersicon esculentum	Tomato	1	3.2-fold smaller
Medicago truncatula	Barrel medic	0.5	6.4-fold smaller
Oryza sativa	Rice	0.466	6.9-fold smaller
Triticum aestivum	Bread wheat	16.5	5-fold larger
Zea mays	Corn	2.365	1.4-fold smaller

The eudicots (such as *Arabidopsis*) diverged from the monocots (such as *O. sativa*) about 200 MYA. Among the eudicots, the rosids and the asterids diverged about 100–150 MYA (Allen, 2002). The rosids include *Arabidopsis, Glycine max* (soybean), and *M. trunculata*. The asterids include *Lycopersicon esculentum* (tomato).

respectively) (Table 16.16). This made it an attractive choice as the first plant genome to be sequenced. The Arabidopsis Genome Initiative (2000) reported the sequence of most (115 Mb) of the genome. There are five chromosomes, predicted to encode 25,498 genes. The *Arabidopsis* genome has an average density of one gene per 4.5 kb.

The estimated number of predicted genes in *Arabidopsis* has increased slightly, following reannotation of the genome by TIGR and by Crowe et al. (2003) (see sidebar). *Arabidopsis* has considerably more genes than *Drosophila* (about 13,000 genes) and *C. elegans* (about 19,000 genes; see below). The larger number of plant genes can be accounted for by a far greater extent of tandem gene duplications and segmental duplications. There is a core of about 11,600 distinct proteins, while the remaining genes are paralogs (The Arabidopsis Genome Initiative, 2000).

A further surprising feature of *Arabidopsis* is that the whole genome may have duplicated twice. Within the genome there are 24 large, duplicated segments of 100 kb or more, spanning 58% of the genome (The Arabidopsis Genome Initiative, 2000). A comparison of tomato genomic DNA with *Arabidopsis* revealed conserved gene content and gene order with four different *Arabidopsis* chromosomes (Ku et al., 2000). The presence of duplicated and triplicated genomic regions suggests that two (or more) large-scale genome duplication events occurred. One event was ancient, while another occurred about 112 MYA. Following whole-genome duplication, gene

The sequencing involved a collaboration between TIGR, MIPS, and The *Arabidopsis* Information Resource (TAIR) (see Table 16.17).

The TIGR *Arabidopsis* database is available at ▸ http://www.tigr.org/tdb/e2k1/ath1/. This includes many resources, such as an annotation database, and a map of segmental duplications in the *Arabidopsis* genome (▸ http://www.tigr.org/tdb/e2k1/ath1/arabGenomeDups.html). A TIGR *Arabidopsis* gene index lists 22,485 gene models ("tentative consensus" sequences) and weaker evidence for an additional 16,000 genes (AtGI Release 9.0, October 2002).

The Complete *Arabidopsis* Transcriptome Micro Array (CATMA) database is online at ▸ http://www.catma.org/ (Crowe et al., 2003).

TABLE 16-17 Genomics Resources for *Arabidopsis thaliana*

Resource	Comment	URL
TAIR	The Arabidopsis Information Resource	▸ http://www.arabidopsis.org/
Arabidopsis thaliana Database	At TIGR	▸ http://www.tigr.org/tdb/e2k1/ath1/
Arabidopsis thaliana Project	At MIPS	▸ http://www.mips.biochem.mpg.de/proj/thal/proj/thal.overview.html
Arabidopsis genome analysis	At Cold Spring Harbor	▸ http://nucleus.cshl.org/protarab/
SeedGenes	Essential genes	▸ http://www.seedgenes.org

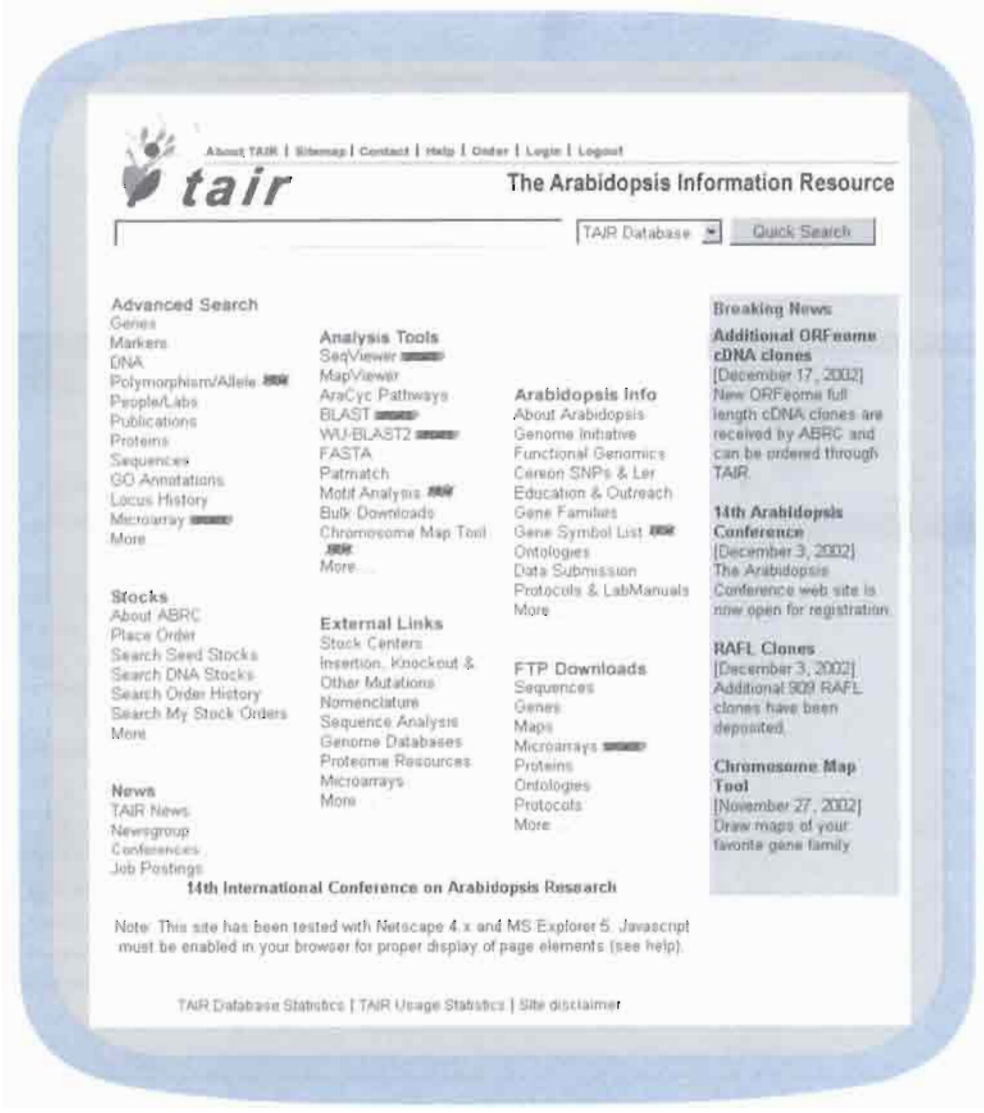


FIGURE 16.24. *The Arabidopsis Information Resource is a comprehensive site for molecular information on* Arabidopsis.

loss occurred frequently. This reduces the amount of gene colinearity observed today and hinders our ability to decipher the nature and timing of past polyploidization events (Simillion et al., 2002).

Several *Arabidopsis* genomics resources are listed in Table 16.17. The most comprehensive site is TAIR, with a wide range of services (Rhee et al., 2003) (Fig. 16.24). This site provides access to genomic DNA sequence from the broadest chromosome-level view to descriptions of single-nucleotide polymorphisms (Fig. 16.25). The format of this site is shared by a variety of genome projects (Box 16.3). Other databases include SeedGenes, which describes essential genes of *Arabidopsis* that give a seed phenotype when disrupted by mutation (Tzafrir et al., 2003).

Large segmental duplications in *Arabidopsis* were identified using MUMmer (see Chapter 14) and tblastx searches (see Chapter 4).

Rice

By some estimates, rice (*O. sativa*) is the staple food for half the human population. The rice genome was the second plant genome to be sequenced. At approximately 430 Mb, this genome size is about one-eighth that of the human genome. Still it

Grasses include rice, wheat, maize, sorghum, barley, sugarcane, millet, oat, and rye. There are over 10,000 species of grasses (Bennetzen and Freeling, 1997).

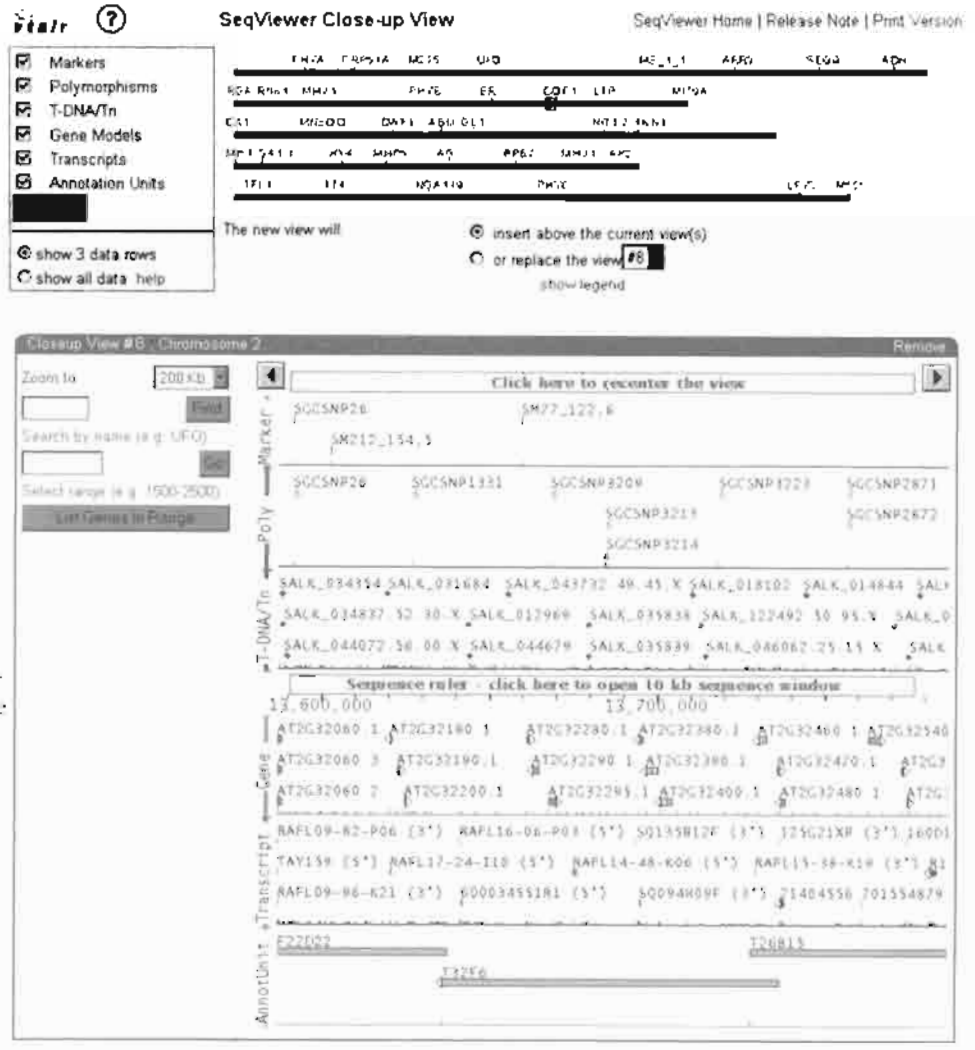


FIGURE 16.25. *The TAIR web browser includes a sequence viewer that allows broad views of the genomic landscape as well as detailed views.*

Cereals are seeds of flowering plants of the grass family (Gramineae, also called Poaceae) that are cultivated for the food value of their grains. Grasses are monocotyledonous plants that range from small, twisted, erect, or creeping annuals to perennials. See the rice glossary at ▸ http://www.riceweb.org/glossary/Terms.htm.

The International Rice Genome Sequencing Project (IRGSP) produced a draft version of the *O. sativa* ssp. *japonica* genome (▸ http://genome.sinica.edu.tw/). Two companies also sequenced this genome (Monsanto at ▸ http://www.rice-research.org and Syngenta at ▸ http://www.tmri.org). The Beijing Genomics Institute led a consortium that generated a draft version of the subspecies *indica* genome (see ▸ http://btn.genomics.org.cn/rice) (Yu et al., 2002). The principal international consortium for rice genomics from Japan has a website (▸ http://rgp.dna.affrc.go.jp/), as does the United States Rice Genome Sequencing effort (▸ http://www.usricegenome.org/).

The TIGR Rice Genome Project database is online at ▸ http://www.tigr.org/tdb/e2k1/osa1/ (Yuan et al., 2003). The MIPS (*O. sativa*) database (MOsDB) is available at ▸ http://mips.gsf.de/proj/rice (Karlowski et al., 2003).

is one of the smallest genomes among the grasses, and rice is studied as a model monocot species.

Four groups generated draft versions of the rice genome (Buell, 2002), including two subspecies. A consortium led by the Beijing Genomics Institute reported a draft sequence of the rice genome (*O. sativa* L. ssp. *indica*) (Yu et al., 2002). Another consortium reported a draft genome sequence of a different rice subspecies, *O. sativa* L. ssp. *japonica*) (Goff et al., 2002). As discussed above (pages 560–562), the annotation of genes and other features is far superior in finished sequence of chromosomes 1 (Sasaki et al., 2002) and 4 (Feng et al., 2002) relative to draft sequence.

Several databases provide comprehensive collections of genomic and other data on rice, such as TIGR and MOsDB. These sites provide extensive genome annotation.

The rice genome (subspecies *indica*) displays an unusual feature of a gradient in GC content. The mean GC content is 43.3%, higher than in *Arabidopsis* (34.9%) or human (41.1%) (Yu et al., 2002). A plot of the number of 500-bp sequences (*y* axis)

BOX 16-3
Databases for Eukaryotic Genomes

The main *Arabidopsis* database, TAIR, uses a database template shared by other major sequencing projects (Table 16.18). We already explored EcoCyc in Chapter 8 and the yeast database SGD in Chapter 15. These databases offer both detailed and extremely broad views of the genomic landscape (Gelbart, 1998). The Genomics Unified Schema (GUS) is another commonly used platform (Table 16.19). Many databases use a distributed annotation system (DAS) that allows a computer server to integrate genomic data from a variety of external computer systems. DAS, written by Lincoln Stein and Robin Dowell, is described at biodas.org (▸ http://www.biodas.org/). It is employed at WormBase, FlyBase, Ensembl, and TIGR sites, among others.

TABLE 16-18 Variety of Databases Employing Template from Generic Model Organism Project (GMOD) (▸ http://www.gmod.org/)

Database	Comment	URL
EcoCyc	Encyclopedia of *Escherichia coli* Genes and Metabolism	▸ http://EcoCyc.org/
FlyBase	*Drosophila* site	▸ http://www.flybase.org/
Mouse Genome Informatics	Main mouse resource	▸ http://www.informatics.jax.org/
Rat Genome Database (RGD)	Rat resource	▸ http://rgd.mcw.edu/
SGD	See Chapter 15	▸ http://genome-www.stanford.edu/Saccharomyces/
TAIR	The *Arabidopsis* Information Resource	▸ http://www.arabidopsis.org/
Wormbase	*C. Elegans* Site	▸ http://www.wormbase.org/

TABLE 16-19 Genomics Unified Schema Platform (▸ http://www.gusdb.org/)

This platform is used for some organism databases

Database	Comment	URL
AllGenes	Human and mouse gene index	▸ http://www.allgenes.org/
EPConDB	Endocrine Pancreas Consortium	▸ http://www.cbil.upenn.edu/EPConDB/
GeneDB	Curated database for *S. pombe, Leishmania major,* and *T. brucei*	▸ http://www.genedb.org/
PlasmoDB	Genomic database for *P. falciparum*	▸ http://plasmodb.org/
RAD	RNA abundance database	▸ http://www.cbil.upenn.edu/RAD2/

versus the percent GC content (x axis) revealed a tail of many GC-rich sequences. These GC-rich regions occurred selectively in rice exons (rather than introns), and at least one exon of extremely high GC content was found in almost every rice gene (Yu et al., 2002). The GC content of the 5′ end of each gene was typically 25% more GC rich than the 3′ end. These unique features of the rice genome present another major challenge for the use of ab initio gene-finding software.

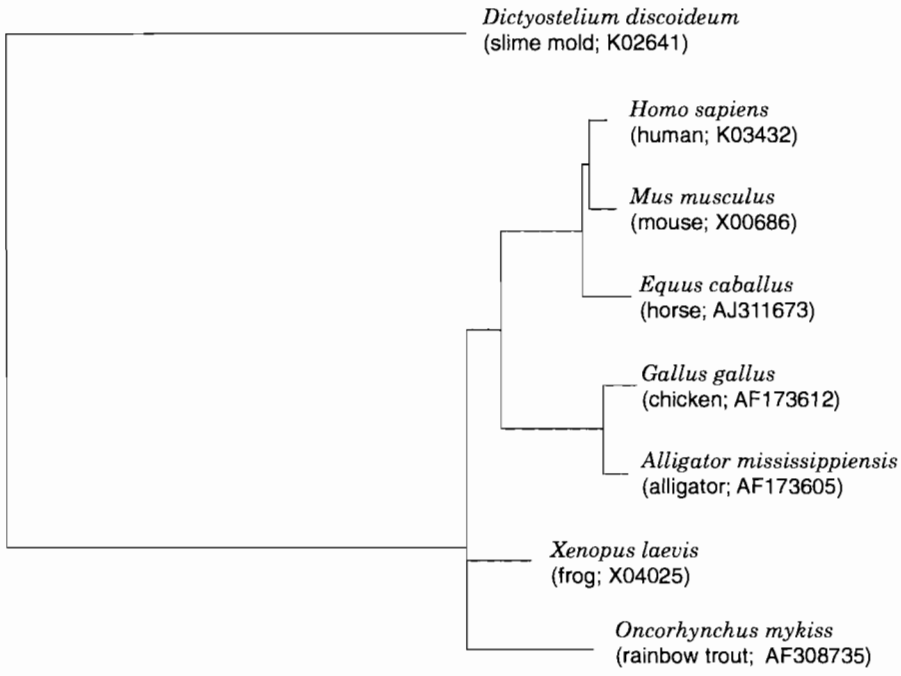


FIGURE 16.26. *The slime mold* D. discoideum *is closely related to the metazoans, as shown in this neighbor-joining tree from 18S ribosomal RNA. This phylogeny is consistent with that shown in Figure 16.1.*

Slime at the Feet of Metazoans

The metazoans are familiar to us as animals, including worms, insects, fish, and mammals. The slime mold *D. discoideum* is a social amoeba that is of great interest as a eukaryote that is an outgroup of the metazoa. This is seen in Figures 16.1 and 16.26, the latter figure depicting a phylogeny based on 18S ribosomal RNA.

Biologists have studied *Dictyostelium* because of its remarkable life cycle. In normal conditions it is a single-celled organism that occupies a niche in soil. Upon conditions of starvation, it emits pulses of cyclic AMP (cAMP), promoting the aggregation of large numbers of amoebae. This results in the formation of an organism having the properties of other multicellular eukaryotes: It differentiates into several cell types, responds to heat and light, and undergoes a developmental profile.

For a description of the *Dictyostelium* genome-sequencing project, see an article by William F. Loomis at ▸ http://dictybase.org/genomeseq.htm.

The *Dictyostelium* genome is 34 Mb, localized on six chromosomes. There are also about 100 copies of a 90-kb palindromic chromosome containing the rRNA genes and a mitochrondrial genome (55 kb). The largest of the main chromosomes, chromosome 2, consists of 8 Mb and was sequenced by an international consortium (Glockner et al., 2002). Because the genome consists of almost 80% AT content—similar to *P. falciparum*—as well as many repetitive DNA sequences, a whole-chromosome shotgun strategy has been adopted. Chromosome 2 has 2799 predicted proteins, suggesting that the entire genome consists of about 11,000 genes. The gene density is high (one gene per 2.6 kb), and there are relatively few introns (1.2 per gene). These introns have an AT content of 87%, while in exons the AT content is 72%. This discrepant compositional bias may represent a mechanism by which introns are spliced out (Glockner et al., 2002).

For websites that offer *Dictyostelium* resources, see Table 16.20.

Introduction to Metazoans

For alternative classification systems, see Cavalier-Smith (1998) and Margulis and Schwartz (1998). For an animal phylogeny based on cytochrome *c* oxidase I see Hebert et al. (2003). Karl Leuckart (1822–1898) first divided the metazoa into six phyla.

The metazoans include most of the animals that are familiar to us. The NCBI taxonomy site classifies five superphyla (Table 16.21). The bilateria are further divided

TABLE 16-20 Genomics Resources for *Dictyostelium discoideum*

Resource	Comment	URL
Dictybase	A centralized source for information about *Dictyostelium* and related organisms	▸http://dictybase.org/
The *Dictyostelium* Genome Sequencing Project	Baylor College of Medicine	▸http://dictygenome.bcm.tmc.edu/
	Institute of Biochemistry I, Cologne	▸http://www.uni-koeln.de/dictyostelium/
	Department of Genome Analysis, IMB Jena	▸http://genome.imb-jena.de/dictyostelium/
	Sanger Institute	▸http://www.sanger.ac.uk/Projects/D_discoideum/
Dictyostelium links	NCBI	▸http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map00?taxid=44689

into Coelomata (including humans and insects), Acoelomata, and Pseudocoelomata (including nematode worms) (Table 16.22). You can see the relationship of nematodes, arthropods (insects), and chordates (including humans) in Figure 16.22.

The Coelomata (or bilaterian animals) are further divided into two major groups. (1) The protostomes include arthropods (insects), annelids, and mollusks. We will survey the first protostome genomes that have been sequenced: the insects *D. melangoaster* and *A. gambiae*. (2) The deuterostomes consist of the echinoderms (such as the sea urchin *Strongylocentrotus purpuratus*), the hemichordates (such as acorn worms), and the chordates (vertebrates as well as cephalochordates and urochordates). These three deuterostome phyla descended from a common ancestor about 550 MYA, the time of the Cambrian explosion. We will discuss a basal member of the chordates, the urochordate sea squirt *Ciona intestinalis*, and we will then examine the vertebrate genomes of a fish (*F. rubripes*) and the mouse.

The bilateria are bilaterally symmetric animals. The coelom is the body cavity of an animal. Acoelomata lack a body cavity. Coelomata have a body cavity that is lined with mesodermal tissue, while Pseudocoelomata such as the nematode *C. elegans* have a body cavity lined with other tissue.

Analysis of a Simple Animal: Nematode *Caenorhabditis elegans*

Caenorhabditis elegans is a free-living soil nematode. It has served as a model organism because it is small (about 1 mm in length), easy to propagate (its life cycle is three days), has an invariant cell lineage that is fully described, and is suitable for many genetic manipulations. Furthermore, it has a variety of complex physiological traits characteristic of higher metazoans such as vertebrates, including an advanced

The phylum Cnidaria is an outgroup to the bilateria, having diverged about 600–750 MYA. Its members include sea anemones, hydras, corals, and jellyfishes. CnidBase organizes genomic and other information on diverse cnidarians (▸http://cnidbase.bu.edu) (Ryan and Finnerty, 2003).

The soma of an adult hermaphrodite worm consists of 959 cells, including 302 cells in the central nervous system.

TABLE 16-21 Metazoan (Animal) Kingdom as Classified at NCBI (▸http://www.ncbi.nlm.nih.gov/Taxonomy/)

Eumetazoan denotes "true metazoan." The bilateria (boldface) include the vertebrates

Superphylum	Phylum	Species Example
Eumetazoa	**Bilateria**	*Homo sapiens*
	Cnidaria (corals, jellyfish)	*Polyorchis penicillatus*
	Ctenophora	*Beroe ovata*
Mesozoa	Orthonectida	*Rhophalura ophiocomae*
	Rhombozoa	*Dicyema misakiense*
Myxozoa	Myxosporea	*Myxobolus cerebralis*
Placozoa	Trichoplax	*Trichoplax adhaerens*
Porifera (sponges)	Calcarea	*Sycon raphanus*
	Demospongiae	*Geodia cydonium*
	Hexactinellida	*Aphrocallistes vastus*

TABLE 16-22 Phylum Bilateria

The first three animal genomes to be sequenced were *C. elegans, Drosophila,* and human (boldface)

Superphylum	Phylum	Species Example
Acoelomata	Platyhelminthes (flatworms)	*Schistosoma mansoni*
Coelomata	***Deuterostomia (e.g. chordates, echinoderms, arrow worms)***	***Homo sapiens***
	Protostomia [e.g. mollusks, leeches, arthropods (insects)]	***Drosophila melanogaster***
Pseudocoelomata	Acanthocephala (thorny-headed worms)	*Moniliformis moniliformis*
	Cycliophora	*Symbion pandora*
	Gastrotricha	*Draculiciteria tesselata*
	Kinorhyncha	*Pycnophyes kielensis*
	Nematoda (roundworms)	***Caenorhabditis elegans***
	Nematomorpha (horsehair worms)	*Gordius aquaticus*
	Rotifera	*Adineta vaga*

Source: Adapted from the classification at NCBI (▸http://www.ncbi.nlm.nih.gov/Taxonomy/).

central nervous system. Many nematodes are parasitic, and an understanding of *C. elegans* biology may lead to treatments for a variety of human diseases.

The 2002 Nobel Prize in Physiology or Medicine was awarded to three researchers who pioneered the use of *C. elegans* as a model organism: Sydney Brenner, H. Robert Horvitz, and John E. Sulston. See ▸http://www.nobel.se/medicine/laureates/2002/.

Another advantage of studying *C. elegans* is that its genome size of 97 Mb is relatively small. This genome was the first of an animal and the first of a multicellular organism to be sequenced [The *C. elegans* Sequencing Consortium (CESC), 1998]. The genome sequencing was based on physical maps of the five autosomes and single X chromosome. The GC content is an unremarkable 36%. It was predicted that there are 19,099 protein-coding genes, with 27% of the genome consisting of exons. About 42% of *C. elegans* proteins have predicted orthologs outside Nematoda, while 34% match only other nematode proteins.

About 300 species of parasitic worms infect human (Cox, 2002). While 20,000 nematode species have been described, it is thought that there may be one million species (Blaxter, 1998, 2003).

The *C. elegans* proteome contains a large number of predicted seven-transmembrane-domain (7TM) receptors of both the chemoreceptor family and rhodopsin family (Table 16.23). This illustrates the principle that new protein functions can emerge following gene duplication (Sonnhammer and Durbin, 1997). It is also notable that many nematode proteins are absent from nonmetazoan species (plants and fungi).

The principal web resource for *C. elegans* is WormBase, a comprehensive database (Harris et al., 2003). WormBase features a variety of data, including:

- Genomic sequence data
- The developmental lineage
- The connectivity of the nervous system
- Mutant phenotypes, genetic markers, and genetic map data
- Gene expression data
- Bibliographic resources

WormBase is available at ▸http://www.wormbase.org.

Caenorhabditis elegans has been the subject of many functional genomics projects (Fields et al., 1999; Kim, 2001; Brooks and Isaac, 2002). Gene expression has been measured using microarrays at six different developmental stages (Hill et al., 2000). About 10,700 open reading frames (56%) were detected in at least one hybridization. This number is comparable to the complement of expressed sequence tags, and the remaining thousands of genes may be expressed in specialized body regions,

TABLE 16-23 Pfam Classification of Ten Most Common *C. elegans* Protein Domains

	Number of Proteins Matched					
InterPro	Worm[a]	Plant[b]	Fly[c]	Human	Yeast[d]	Name
IPR000168	638	0	0	0	0	Nematode 7TM chemoreceptor (probably olfactory)
IPR000719	512	1046	265	688	118	Protein kinase
IPR000276	403	6	86	826	0	Rhodopsin-like GPCR superfamily
IPR000379	136	209	139	112	37	Esterase/lipase/thioesterase, active site
IPR0007114	127	83	114	100	67	Major facilitator superfamily (MFS)
IPR000387	106	25	37	139	10	Tyrosine-specific protein phosphatase and dual-specificity protein phosphatase
IPR000324	104	0	17	18	0	Vitamin D receptor
IPR005821	100	30	58	149	3	Ion transport protein
IPR005828	96	90	103	75	53	General substrate transporter
IPR006201	95	0	20	43	0	Neurotransmitter-gated ion channel

Abbreviations: 7TM, seven transmembrane domain receptor; GPCR, G-protein-coupled receptor.
[a] *C. elegans.*
[b] *A. thaliana.*
[c] *D. Melanogaster*
[d] *S. cerevisiae.*
Source: Adapted from the European Bioinformatics Institute (▸http://www.ebi.ac.uk/proteome/).

developmental stages, or physiological conditions. In another approach to defining gene function, Kamath et al. (2003) inhibited the function of 86% of the >19,000 predicted *C. elegans* genes using RNA interference (RNAi). They identified mutant phenotype 1 for 1,722 genes.

First Insect Genome: *Drosophila melanogaster*

The arthropods may be the most successful set of eukaryotes on the planet in terms of the number of species. They include the Chelicerates—such as the scorpions, spiders, and mites—and the Mandibulata, animals with modified appendages (mandibles) such as the insects (Table 16.24).

The fruit fly *D. melanogaster* has been an important model organism in biology for a century (Rubin and Lewis, 2000). The fly is ideal for studies of genetics because of its short life cycle (two weeks), varied phenotypes (from changes in eye

About 1 million arthropod species have been described, but there are an estimated 3–30 million species (Blaxter, 2003).

Thomas Hunt Morgan was awarded a Nobel Prize in 1933 "for his discoveries concerning the role played by the chromosome in heredity." See ▸http://www.nobel.se/medicine/laureates/1933/. In 1995, Edward B. Lewis, Christiane Nüsslein-Volhard, and Eric F. Wieschaus shared a Nobel Prize "for their discoveries concerning the genetic control of early embryonic development." These studies concerned *Drosophila* development (▸http://www.nobel.se/medicine/laureates/1995/).

TABLE 16-24 Arthropods (Phylum Arthropoda) as Classified at NCBI (▸http://www.ncbi.nlm.nih.gov/Taxonomy/)

Arthropods are invertebrate protostomes (see Table 16.22). Pancrustacea (boldface) is further divided into the superclasses Crustacea (crustaceans) and Hexapoda (insects). Insecta includes *D. melanogaster* and *A. gambiae*

Subphylum	Class
Chelicerata	Arachnida (mites, ticks, spiders) Merostomata (horseshoe crabs) Pycnogonida (sea spiders)
Mandibulata	Myriapoda (centipedes) **Pancrustacea (crustaceans, insects)**

color to changes in behavior, development, or morphology), and large polytene chromosomes that are easily observed under a microscope.

The *Drosophila* genome was sequenced based in large part upon the whole-genome shotgun sequencing strategy (Adams et al., 2000). Prior to this effort, the whole-genome shotgun strategy had only been applied to far smaller genomes, and thus this success represented a significant breakthrough. The 180-Mb genome is organized into an X chromosome (numbered 1), two principal autosomes (numbered 2 and 3), a very small third autosome (numbered 4; about 1 Mb in length), and a Y chromosome. Approximately one-third of the genome contains heterochromatin (mostly simple sequence repeats as well as transposable elements and tandem arrays of rRNA genes). This heterochromatin is distributed around the centromeres and across the length of the Y chromosome. The transition zones at the boundary of heterochromatin and euchromatin contained many protein-coding genes that were previously unknown.

The *Drosophila* genome was sequenced through a collaborative effort that included Celera Genomics, the Berkeley *Drosophila* Genome Project (BDGP; ▸http://www.fruitfly.org), and the European *Drosophila* Genome Project (EDGP) (Adams et al., 2000). As part of the effort, three million random genomic fragments of ≈500 bp were sequenced.

The initial annotation of the *Drosophila* genome described 17,464 genes predicted with Genscan and 13,189 genes predicted with the Genie algorithm (Reese et al., 2000; Adams et al., 2000). The authors believed that Genscan overestimated the true number of genes. Subsequently, the genome sequence was finished to close gaps and to improve sequence quality. This resulted in releases 2 and 3 of the genome sequence (Celniker et al., 2002). In release 3, the total number of predicted protein-coding genes was essentially unchanged (13,676) relative to the initial genome annotation (Misra et al., 2002). However, the estimated total number of unique exons increased substantially. Furthermore, there were changes to the models for 85% of the transcripts and for 45% of the predicted proteins. The improved annotation can be attributed to the availability of more expressed sequence tags and complete cDNAs that can be aligned to genomic DNA. These studies have clarified the extent of untranslated regions, which are difficult to assign using ab initio gene-finding algorithms.

The principal database for *D. melanogaster* (and for other species of the family Drosophilidae) is FlyBase (▸http://flybase.bio.indiana.edu/) (The FlyBase Consortium, 2003).

According to release 3, 13,379 protein-coding genes are in the euchromatin and 297 protein-coding genes are in the heterochromatin (Misra et al., 2002).

While one task of genome annotation is the description of protein-coding genes, another task is to assign function to those genes. This can be approached by assessing the extent to which predicted proteins have identifiable orthologs. Rubin et al. (2000) systematically blastp searched the proteomes of the fly, *C. elegans,* and *S. cerevisiae*. They drew several conclusions:

- The "core proteomes" of these organisms represent the set of unique proteins, excluding paralogs. These sizes are 4383 proteins (yeast), 8065 proteins (fly), and 9453 proteins (worm). Thus, despite the fact that *S. cerevisiae* is unicellular while fly and worm are multicellular, the core proteomes are only twofold different.
- About 30% of the fly genes have orthologs in worm; 20% have an ortholog in both worm and yeast. Such proteins may be in common to all eukaryotic cells.
- Half of the fly proteins have a mammalian homolog (at an *E* value cutoff below 10^{-10}), consistent with a model in which the fly is more closely related to humans than is the worm (see Fig. 16.22).
- A substantial number of *Drosophila* proteins are not significantly related to proteins from yeast, worm, or mammals.
- The fly and worm each have about 2200 multidomain proteins. However, yeast has only 672. Proteomic analyses such as these may elucidate the

Rubin et al. (2000) defined orthologs as having significant similarity (based on *E* values) over at least 80% of the query protein length. This leads to an underestimate of the number of orthologs because some matched regions are small (see e.g., Fig. 8.3).

TABLE 16-25 Human Diseases Borne by Mosquitoes

Disease	Mosquito Species	Number of Cases
Malaria	*Anopheles gambiae*	500 million
Dengue	*Aedes aegypti*	50 million per year
Lymphatic filariasis	*Culex quinquefasciatus, Anopheles gambiae*	120 million
Yellow fever	*Aedes aegypti*	200,000 per year
West Nile virus disease	*Culex tarsalis, Culex pipiens,* other	≈3400 per year

Source: Adapted from Budiansky (2002) and Holt et al. (2002). West Nile virus disease data are for January 1 to November 30, 2002 in the United States (Centers for Disease Control and Prevention, 2002).

molecular basis of phenotypic differences between organisms. For example, in contrast to yeast, the fly and worm have many proteins with extracellular domains having roles in cell–cell contact and cell–substrate contact.

Second Insect Genome: *Anopheles gambiae*

The mosquito *A. gambiae* is most well known as the malaria vector that carries the protozoan parasite *P. falciparum* (as well as *P. vivax, P. malariae,* and *P. ovale*). Mosquitoes are responsible for a variety of human diseases, although most of these (except West Nile) are generally restricted to the tropics (Table 16.25).

Holt et al. (2002) reported the genomic sequence of a strain of *A. gambiae* using the whole-genome sequencing strategy. The genome is 278 Mb arranged in an X chromosome (numbered 1) and two autosomes (numbered 2 and 3). A particular challenge in sequencing this genome is the high degree of genetic variation, as manifested in "single-nucleotide discrepancies." Thus there is a mosaic genome structure caused by two haplotypes of approximately equal abundance. In contrast, the *D. melanogaster* and *M. musculus* genomes are entirely homozygous.

You can view and explore the *A. gambiae* genome at the Ensembl genome browser. Annotation by the Ensembl pipeline (Chapter 12) and Celera suggests the existence of 13,683 genes. As with all eukaryotic genome projects, the *A. gambiae* annotation is known to contain many incomplete or incorrect gene assignments (Holt et al., 2002).

The *A. gambiae* genome is about 278 Mb, or more than twice the size of *Drosophila*. This difference in genome size is accounted for by intergenic DNA, and *Drosophila* appears to have undergone a genome size reduction relative to *Anopheles* species and to other *Drosophila* species (Holt et al., 2002). *Anopheles gambiae* and *D. melanogaster* diverged about 250 MYA (Zdobnov et al., 2002). Almost half the genes in these genomes are orthologs, with an average amino acid sequence identity of 56%. By comparison, the lineage leading to modern humans and pufferfish (see below) diverged 450 MYA, but proteins from those two species share a comparable sequence identity (61%). Thus, insect proteins diverge at a faster rate than vertebrate proteins. An outstanding problem is to understand the ability of *Anopheles* to feed on human blood selectively and to identify therapeutic targets. For this effort, it is important to identify arthropod-specific and *Anopheles*-specific genes (Zdobnov et al., 2002).

Road to Vertebrates: *Ciona intestinalis*

The vertebrates include fish, amphibians, reptiles, birds, and mammals. All these creatures have in common a segmented spinal column. From where did the

A haplotype is a combination of alleles of closely linked loci that are found in a single chromosome and tend to be inherited together.

The Ensembl genome browser for the mosquito is available at ▸ http://www.ensembl.org/Anopheles_gambiae/. The main Ensembl page (▸ http://www.ensembl.org) includes links to the human, mouse, rat, fruitfly, nematode, zebrafish, and fugu genomes. We will explore this important site for the human genome in Chapter 17.

We described the *Drosophila* Down syndrome cell adhesion molecule (DSCAM) in Chapter 6, a gene that potentially encodes up to 38,000 distinct proteins through alternative splicing (NP_523649). The *A. gambiae* ortholog appears to share the same potential for massive alternative splicing (Zdobnov et al., 2002). See GenBank protein accession EAA05472.

One of the smallest chordate genomes is that of *Oikopleura dioica,* estimated to be between 51 and 72 Mb (Seo et al., 2001). This ocean-dwelling animal may have about 15,000 genes.

vertebrates originate? Vertebrates are members of the chordates, animals having a notochord. The sea squirt *C. intestinalis* is a urochordate (also called tunicate), one of the subphyla of chordates but not a vertebrate. *Ciona* is a hermaphroditic invertebrate that offers us a window on the transition to vertebrates (Holland, 2002).

Dehal et al. (2002) produced a draft sequence of the *C. intestinalis* genome by the whole-genome shotgun strategy. At 160 Mb it is about 12 times larger than typical fungal genomes and 20 times smaller than the human genome. There are 15,852 predicted genes organized on 14 chromosomes. Most of these predicted genes are supported by evidence from expressed sequence tags.

The availability of the *Ciona* genome sequence allows a comparison with protostomes and other deuterostomes and supports its position as related to an ancestral chordate (Dehal et al., 2002). Almost 60% of *Ciona* genes have protostome orthologs; these presumably represent ancient bilaterian genes. Several hundred genes have invertebrate but not vertebrate homologs, such as the oxygen carrier hemocyanin.

There are 2570 *Ciona* genes (one-sixth) that have orthologs in vertebrates but none in protostomes; these genes arose in the deuterostome lineage before the last common ancestor diverged into vertebrates, cephalochordates, and urochordates (e.g., *Ciona*). There are 3399 *Ciona* genes (one-fifth) that have no identifiable homolog in vertebrates or invertebrates and thus may be tunicate-specific genes that evolved after the divergence of the urochordate lineage.

Ciona has genes involved in processes such as apoptosis (programmed cell death), thyroid function, neural function, and muscle action. This provides an opportunity for comparative analyses of fundamentally important genes within the chordate lineage. For example, nerves communicate with muscles by releasing the neurotransmitter acetylcholine from synaptic vesicles in presynaptic nerve terminals. This transmitter diffuses across the synapse (a gap between cells) to bind and activate postsynaptic receptors. *Ciona* has genes encoding proteins that function in neurotransmission, including a transferase enzyme that synthesizes acetycholine, an acetycholine transporter that pumps the neurotransmitter into vesicles, synaptic vesicle proteins, and neurotransmitter receptors.

The Department of Energy Joint Genome Institute operates the *C. intenstinalis* genome home page (▸http://genome.jgi-psf.org/ciona4/ciona4.home.html). The GenBank accession number for the genome is AABS00000000, and you can find a *Ciona* BLAST server through the NCBI Genomes page of eukaryotic projects. TIGR has a *Ciona* gene index (▸http://www.tigr.org/tdb/tgi/cingi/). The Ghost database, a *Ciona* EST project that includes a BLAST server and gene expression data, is available at ▸http://ghost.zool.kyoto-u.ac.jp/indexr1.html.

A *Ciona* protein (BAB85859) has 47% identity to human choline acetyltransferase (NP_065574). A *Ciona* gene (accession AB071998) encodes a protein with 56% identity to a human vesicular acetylcholine transporter (NP_003046). Note that many such genes also function in neurotransmission in invertebrates.

Vertebrate Genome of a Fish, *Fugu rubripes*

The Japanese pufferfish *F. rubripes* is a vertebrate having a remarkably compact genome. This teleost fish has a genome size of 365 Mb, about one-tenth the size of the human genome (Aparicio et al., 2002). However, *Fugu* and humans—two species that diverged about 450 MYA—have comparable numbers of predicted protein-coding genes.

There are several reasons that the *Fugu* genome is relatively compact (Aparicio et al., 2002):

- Only 2.7% of the *Fugu* genome consists of interspersed repeats, based on analyses with RepeatMasker. This contrasts with 45% interspersed repeats in the human genome (Chapter 17). Still, every known class of eukaryotic transposable elements is represented in *Fugu*. The most common *Fugu* repeat is the LINE-like element *Maui* (6400 copies), while in humans there are over one million copies of the most common repeat, *Alu*.
- Introns are relatively short. Seventy-five percent of *Fugu* introns are <425 bp in length, while in humans 75% of introns are <2609 bp. In *Fugu*, about 500 introns have a length greater than 10 kb, while in humans more than 12,000 introns are greater than 10 kb.

Fugu rubripes is also called *Takifugu rubripes*. The International *Fugu* Genome Consortium was responsible for the sequencing of its genome. A major gateway to *Fugu* resources is at the U.S. Department of Energy Joint Genome Institute site (▸http://genome.jgi-psf.org/fugu6/fugu6.info.html).

The Fugu Browser is at ▸http://www.ensembl.org/Fugu_rubripes/. Produced by the Wellcome Trust Sanger Institute and the European Bioinformatics Institute, it is a major portal to this genome and others.

- Gene loci occupy about 108 Mb of the total euchromatic DNA (320 Mb). This represents about one-third of the genome, a far higher fraction than in mouse or human.

For a database that collates data on over 27,000 species of fish, see ▸http://www.fishbase.org/.

Analysis of Mammalian Genome: Mouse

The sequencing and analysis of the mouse genome represents a landmark in the history of biology. Following the human, the mouse is the second mammal to have its genome sequenced. The mouse is an excellent model for understanding human biology:

Two other fish genome projects are for the zebrafish, *Danio rerio,* and the freshwater fish Medaka (*Oryzias latipes*). See the Zebrafish Information Network at ▸http://zfin.org (Sprague et al., 2003) and the NCBI site ▸http://www.ncbi.nlm.nih.gov/genome/guide/D_rerio.html. See also the Medaka Genome Initiative (▸http://www.dsp.jst.go.jp/MGI) and the Medaka Expression Pattern Database (▸http://medaka.dsp.jst.go.jp/MEPD) (Henrich et al., 2003).

- Remarkably, although these two organisms diverged about 75 MYA, only about 300 of the 30,000 annotated genes in the mouse genome have no counterpart in the human genome.
- In addition to sharing thousands of orthologous protein-coding genes, the mouse and human genomes have large tracts of homologous non-protein-coding DNA. These conserved sequences could give insight into regulatory regions of the genome or noncoding genes (Hardison et al., 1997; Dermitzakis et al., 2002; Dehal et al., 2001).
- The mouse and human share many physiological features. Thus, mice make an important model for hundreds (or thousands) of human diseases, from infectious diseases to complex disorders.
- There are over 1000 mouse strains having spontaneous mutations. Mutations can be introduced into the mouse through random mutagenesis approaches such as chemical mutagenesis or radiation treatment. Mutations and other genetic modifications can also be introduced through directed approaches such as transgenic, knock-out, and knock-in technologies.

Two groups independently sequenced the mouse genome: the Mouse Genome Sequencing Consortium (Waterston et al., 2002) and Celera Genomics (unpublished). These versions of the genome were directly compared by Xuan et al. (2002). They selected over 8300 mouse entries having RefSeq accession numbers as queries and used BLAT (Chapter 5) to compare the coverage and accuracy of the two assemblies. Most mRNAs were matched with both assemblies.

The comparison of the mouse genomes by Xuan et al. (2002), from the laboratory of Michael Zhang, includes a mouse and human genome browser (▸http://143.48.7.130/cgi-bin/gbrowse?source=cse). The Celera effort involved the sequencing of five mouse strains: A/J2, DBA2/J, 129X1/SvJ, 129S1/SvlmJ, and C57BL/6J (see ▸http://www.celera.com).

We will focus our discussion on the public consortium data because they are freely available. Waterston et al. (2002) sequenced the genome of a female mouse of the B6 strain. The sequencing strategy entailed a combined whole-genome shotgun approach (with sevenfold coverage) and a hierarchical shotgun approach (with sequencing of BAC clones that were physically mapped to chromosomes). The assembly covers most of the mouse genome. Of the RefSeq cDNAs, 99.3% could be aligned to the genomic sequence. Also, the Waterston et al. (2002) assembly closely matches an independent draft sequence of mouse chromosome 16 (Mural et al., 2002).

Waterston et al. (2002) described 11 main conclusions of the mouse genome-sequencing project:

1. The total length of the euchromatic mouse genome is 2.5 Gb in size, about 14% smaller than the human genome (2.9 Gb). In contrast to other, more compact genomes we have discussed in this chapter, the mouse genome (like the human genome) averages about one gene every 100,000 bp of genomic DNA. The GC content is comparable, with mean values of 42% (mouse) versus 41% (human). There are 15,500 CpG islands, about half the number observed in humans (see Chapter 17).

The GenBank accession number of the mouse genome is CAAA01000000. It is accessible through the three main human genome sites discussed in Chapter 17: ▸http://www.ensembl.org/Mus_musculus/, ▸http://genome.ucsc.edu/, and ▸http://www.ncbi.nlm.nih.gov.

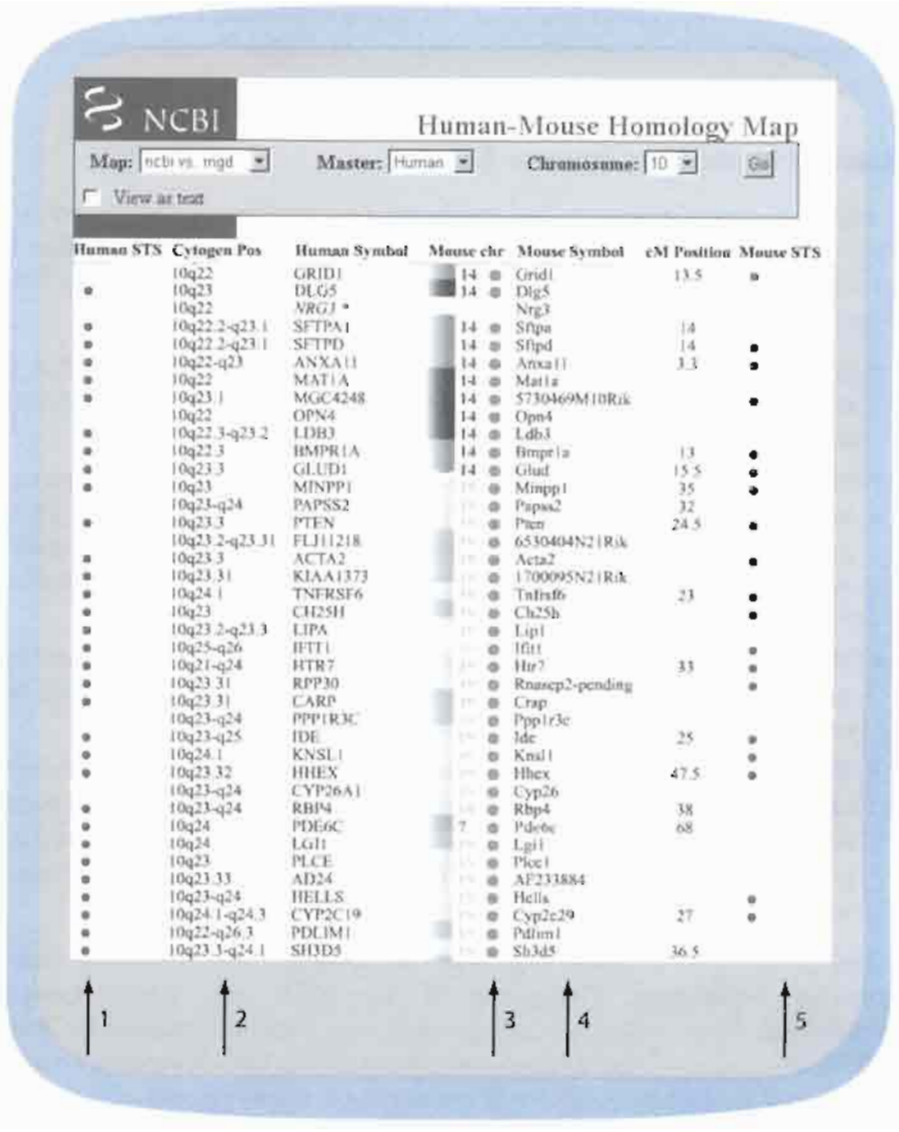


FIGURE 16.27. The NCBI human–mouse homology map shows several different conserved synteny maps. Either human or mouse can be selected as the master; in this case, human chromosome 10 is the master (reference) chromosome. The ouput includes human sequence tag sites (arrow 1), clickable links to LocusLink entries for human (arrow 2) or mouse (arrow 4), a link to the BLAST 2 Sequences pairwise alignment of the two orthologs (arrow 3), and a link to the mouse Map Viewer (arrow 5).

Sequencing a female assured equal coverage of the X chromosome and each autosome. The Y chromosome in both human and mouse is small and contains highly repetitive DNA elements.

You can access the human/mouse homology map from the main page of NCBI or directly at ▸http://www.ncbi.nlm.nih.gov/Homology/.

2. Over 90% of the mouse and human genomes can be aligned into conserved synteny regions. After the divergence of mouse and human about 75 MYA, chromosomal DNA was shuffled in each species. However, large regions of DNA obviously correspond. As an example of how to visualize this, the NCBI offers a human/mouse homology viewer. By selecting human chromosome 10, you can see a map of human genes (including retinol-binding protein) (Fig. 16.27) along with orthologous mouse genes from chromosomes 14, 19, and 7. This conserved synteny map includes links to LocusLink and the human and mouse MapViewers.
3. About 40% of the human genome can be aligned to the mouse genome at the nucleotide level. This represents most of the orthologous sequence shared by these genomes. For 12,845 orthologous gene pairs (see Pitfalls section below), 70.1% of the corresponding amino acid residues were identical.
4. The neutral substitution rate in each genome can be estimated by comparing thousands of repetitive DNA elements to the inferred ancestral

consensus sequence. The average substitution rate is 0.17 per site in humans and 0.34 per site in mouse. The mouse genome also shows a twofold higher rate of acquisition of small (less than 50-bp) insertions and deletions.

5. The proportion of small (50–100-bp) segments in the mammalian genome that is under purifying selection is about 5%. This is estimated by comparing the neutral rate to the extent of sequence conservation in the genome. Since this 5% value is greater than the proportion of protein-coding genes in the genome, genomic regions that do not code for genes must be selected for, such as regulatory elements. Regulatory regions such as those that control liver-specific and muscle-specific expression were conserved between mouse and human to an extent greater than regions of neutral DNA, although less than regions that are protein coding.

The mouse sequencing consortium (Waterston et al., 2002. p. 526) defined a syntenic segment as "a maximal region in which a series of landmarks occur in the same order on a single chromosome in both species." They identified 558,000 orthologous and highly conserved landmarks in the mouse assembly, comprising 7.5% of the mouse assembly.

6. The mammalian genome is evolving in a nonuniform manner, with variation in the rates of sequence divergence across the genome. The neutral substitution rate varied across all chromosomes (and was lowest on the X chromosome), with a higher substitution rate associated with extremes of GC content.

Nadeau and Taylor (1984) estimated that there are about 180 conserved synteny regions of the mouse and human genomes. Waterston et al. (2002) provided evidence for 342 such regions, each greater than 300 kb in size.

7. The mouse and human genomes each contain about 30,000 protein-coding genes. About 80% of mouse genes have a single identifiable human ortholog. Less than 1% of human genes have no identifiable ortholog in the mouse, and vice versa. The sequencing effort revealed the existence of 9000 previously unknown mouse genes as well as 1200 new human genes.
8. Dozens of local gene family expansions have occurred in the mouse genome, such as the olfactory receptor gene family. About 20% of this family are pseudogenes in mouse, suggesting a dynamic interplay between gene expansion and gene deletion. The lipocalins also underwent a mouse lineage-specific expansion. For example, the mouse X chromosome contains a cluster of genes related to odorant-binding protein that are absent in humans. Such expansions may account in part for the physiological differences between primates and rodents in terms of reproductive processes.

The Mouse Genome Sequencing Consortium version of the mouse genome is available at ▸http://www.ensembl.org/Mus_musculus/. This is described at ▸http://www.sanger.ac.uk/Info/Press/2002/021205.shtml. The Celera Genomics version is described at ▸http://www.celera.com.

9. Particular proteins evolve at a rapid rate in mammals. For example, genes involved in the immune response appear to be under positive selection, which drives their evolution.
10. Similar types of repetitive DNA sequences are found in both human and mouse. We will discuss human repetitive sequences in Chapter 17.
11. The public consortium described 80,000 single-nucleotide polymorphisms (SNPs). We will discuss SNPs in Chapter 18.

The most comprehensive mouse resource on the World Wide Web is the Mouse Genome Informatics (MGI) database and its associated sites (see sidebar) (Blake et al., 2003) (Fig. 16.28). A search for *RBP4* results in a detailed report on the gene, including its map position, phenotypes, and expression profile (Fig. 16.29). Additional mouse and rat resources are presented in Tables 16.26 and 16.27.

MGI is available at ▸http://www.informatics.jax.org and is operated by The Jackson Laboratory (▸http://www.jax.org). MGI has multiple components, including the Mouse Genome Database (MGD), the Gene Expression Database (GDX), the Mouse Genome Sequencing (MGS) project, and the Mouse Tumor Biology (MTB) database (▸http://www.informatics.jax.org/mtb).

Primate Genomes

How did humans evolve from other primates? What features of the human genome account for our distinct traits, such as language and higher cognitive skills? A comparison of several primate genomes may elucidate the molecular basis of our unique

FIGURE 16.28. *The Mouse Genome Informatics website from Jackson Laboratory is a principal online resource for the mouse. It includes dozens of essential mouse genomics databases.*

traits—or, depending on one's perspective, such a comparison may highlight how closely similar we are to the great apes at a genetic level.

For an overview of primates, we can begin by making a phylogenetic tree with lysozyme protein sequences (Fig. 16.30). The chimpanzee (*Pan troglodytes*) and the bonobo (pygmy chimpanzee, *Pan paniscus*) are the two species most closely related to humans. These three species diverged from a common ancestor 5.4 ± 1.1 MYA, based on analyses of 36 nuclear genes (Stauffer et al., 2001). Our next closest species is the gorilla, which diverged an estimated 6.4 ± 1.5 MYA. Next in the branching order are the orangutan (11.3 ± 1.3 MYA) and the gibbon (14.9 ± 2.0 MYA) (Stauffer et al., 2001). The hominoids diverged from the Old World monkeys (e.g., the macaque and baboon) 23 MYA, close to the age of the earliest extant hominoid fossils. New World monkeys (such as the tamarin) are even more distantly related.

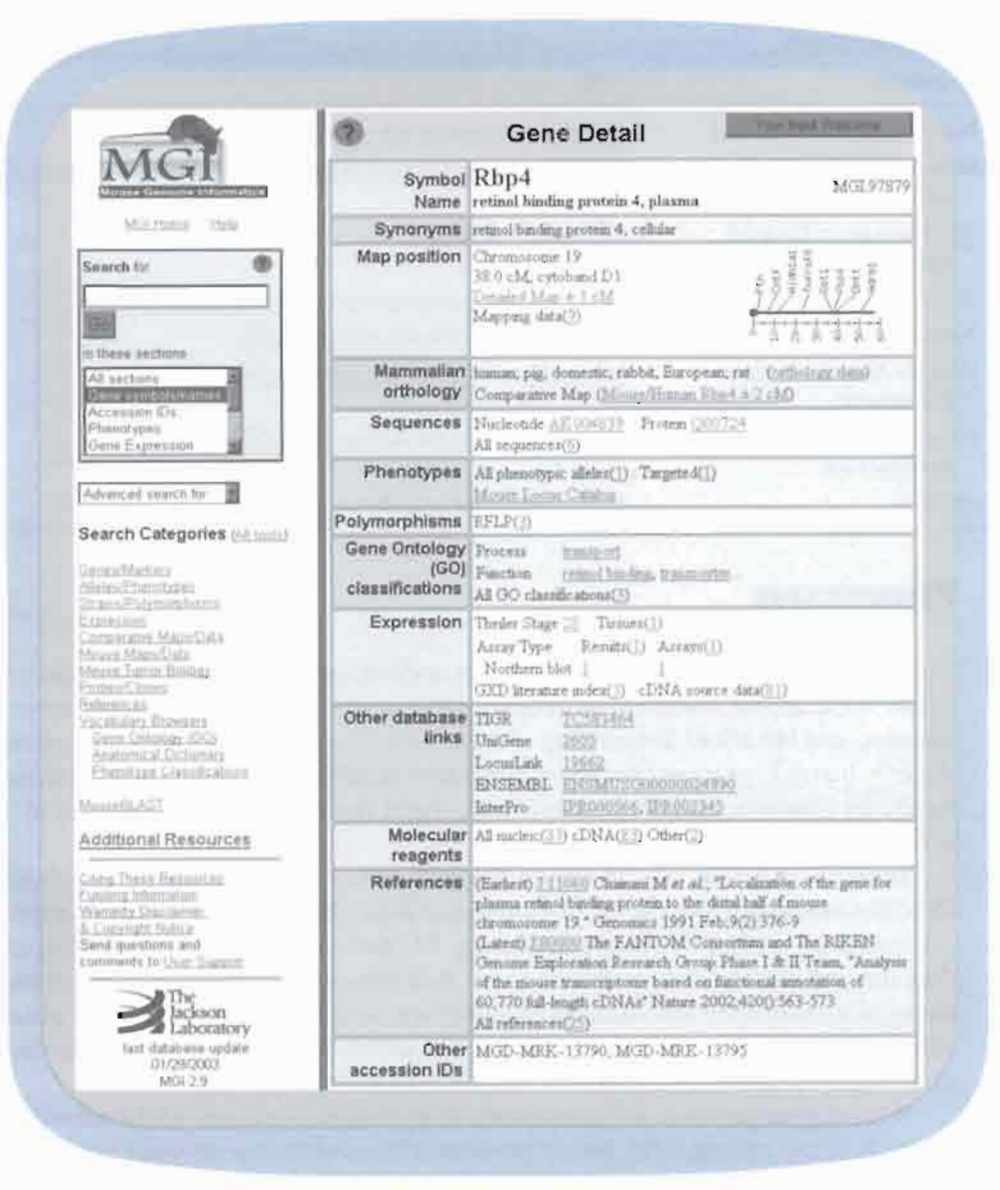


FIGURE 16.29. *The Mouse Genome Informatics report on RBP provides links to data on the map position, putative orthologs, DNA and protein sequence, polymorphisms, gene ontology classifications (see Chapter 8), expression profiles, literature references, and other data.*

Hundreds of protein and nucleotide sequences are currently available for a variety of nonhuman primates in GenBank. Large-scale genome-sequencing projects have begun for the common chimpanzee (*P. troglodytes*). Analysis of the chimpanzee sequence will illuminate many facets of human evolution, genetics, and disease (Reich et al., n.d.; McConkey and Varki, 2000). Human and chimpanzee DNA sequences are above 98% identical.

Other primate genomes under consideration are the rhesus macaque monkey (*Macaca mulatta*) and the olive baboon (*Papio hamadryas anubis*). While these are more distantly related species, they serve as research models more commonly than do chimpanzees, and more data are available on their physiology. Primate resources are listed in Table 16.28.

The African great apes are the chimpanzees and gorillas. For a discussion of earlier primate evolution, including our divergence from tree shrews, bats, and flying lemurs, see Sargis (2002).

TABLE 16-26 Genomics Resources for Mouse, *Mus musculus*

Resource	Comment	URL
Mouse genome project	Baylor College of Medicine	http://www.hgsc.bcm.tmc.edu/projects/mouse/
Mouse cytogenetic maps	Mammalian Genetics Unit, Harwell, United Kingdom	http://www.mgu.har.mrc.ac.uk/anomaly/anomaly-intro.html
TBASE	The Transgenic/Targeted Mutation Database	www.jax.org/tbase
Database of Gene Knockouts	At bioscience.org	http://www.bioscience.org/knockout/knochome.htm
Mouse Genetics	An online book by Lee Silver	http://www.informatics.jax.org/silver/
Mouse mapping	Genome Sciences Centre (Vancouver)	http://www.bcgsc.ca/lab/mapping/mouse
Trans-NIH Mouse Initiative	Comprehensive NIH mouse genomics site	http://www.nih.gov/science/models/mouse/
TIGR Mouse Gene Index	At TIGR	http://www.tigr.org/tdb/mgi/

PERSPECTIVE

One of the broadest goals of biology is to understand the nature of each species of life: What are the mechanisms of development, metabolism, homeostasis, reproduction, and behavior? Sequencing of a genome does not answer these questions directly. Instead, we must first try to annotate the genome sequence in order to estimate its contents, and then we try to interpret the function of these parts in a variety of physiological processes.

In the near future, we can expect the genomes of representative species from all major eukaryotic divisions to become available. This will have dramatic implications for all aspects of eukaryotic biology. For pathogenic organisms, it is hoped that the genome sequence will lead to an understanding of their cellular mechanisms of toxicity, their mechanisms of host immune system evasion, and their pharmacological response to drug treatments. For studies of evolution, we will further understand mutation and selection, the forces that shape genome evolution.

As complete genomes are sequenced, we are becoming aware of the nature of noncoding and coding DNA. Major portions of the eukaryotic genomic landscape are occupied by repetitive DNA, including transposable elements. The number of protein-coding genes varies from about 6000 in fungi to tens of thousands in plants and mammals. Many of these protein-coding genes are paralogous within each species, such that the "core proteome" size is likely to be on the order of 10,000 genes for many eukaryotes. New proteins are invented in evolution through expansions

TABLE 16-27 Genomics Resources for Rat, *Rattus norvegicus*

Resource	Comment	URL
Rat Genome Database	Key rat genomics site	http://rgd.mcw.edu
Ratmap	Rat Genome Database	http://ratmap.gen.gu.se/
NIH Rat Genomics and Genetics	Main NIH rat site	http://www.nih.gov/science/models/rat/
Rat Genome Resources	Central NCBI rat site	http://www.ncbi.nlm.nih.gov/genome/guide/rat/
TIGR Rat Gene Index	At TIGR	http://www.tigr.org/tdb/tgi/rgi/

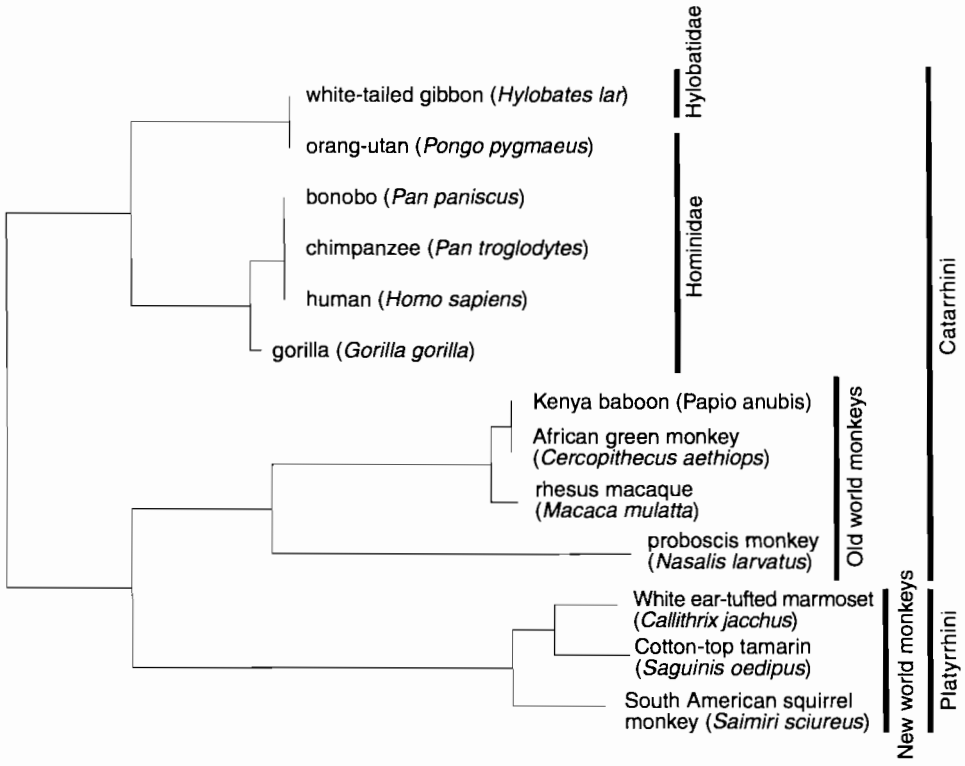


FIGURE 16.30. *A neighbor-joining tree representing primate phylogeny based on lysozyme protein sequences. These sequences were aligned using ClustalW and displayed as a neighbor-joining tree. The accession numbers are as follows: gibbon (P79180), orangutan (P79180), bonobo (AAB41214), chimpanzee (AAB41209), human (P00695), gorilla (P79179), Kenya baboon (P00696), African green monkey (P00696), rhesus macaque (P30201), proboscis monkey (P79811), marmoset (P79158), tamarin (P79268), and South American squirrel monkey (P79294).*

of gene families or through the use of novel combinations of DNA encoding protein domains.

PITFALLS

An urgent need in genomics research is the continued development of algorithms to find protein-coding genes, noncoding RNAs, repetitive sequences, duplicated blocks of sequence within genomes, and conserved syntenic regions shared between genomes. We may then characterize gene function in different developmental stages, body regions, and physiological states. Through these approaches we may generate and test hypotheses about the function, evolution, and biological adaptations of eukaryotes. Thus we may extract meaning from the genomic data.

TABLE 16-28 Genomics Resources for Nonhuman Primates

Resource	Comment	URL
Chimpanzee Genome Project	*Pan troglodytes* at Baylor College of Medicine	http://hgsc.bcm.tmc.edu/projects/chimpanzee/
Project Silver	National Institute of Genetics (Japan)	http://sayer.lab.nig.ac.jp/~silver/
Primate Cytogenetics Network	Diploid numbers, karyotypes, ideograms	http://staff.washington.edu/timk/cyto/

We are now in the earliest years of the field of genomics. Many new lessons are emerging:

- Draft versions of genome sequences are extremely useful resources, but gene annotation often improves dramatically as a sequence becomes finished.
- It is extraordinarily difficult to predict the presence of protein-coding genes in genomic DNA. This is especially true in the absence of complementary experimental data on gene expression, such as expressed sequence tag information.
- We know little about the nature of noncoding RNA molecules.
- Large portions of eukaryotic genomes consist of repetitive DNA elements.
- Comparative genomics is extraordinarily useful in defining the features of each eukaryotic genome.

Most publications describing genomes (both eukaryotic and prokaryotic) define orthologs as descended by speciation from a single gene in a common ancestor. Typically, the predicted proteins from an organism are searched by BLAST against the complete proteome of other species using an E value cutoff such as 10^{-4}. However, two orthologous proteins could have species-specific functions.

Web Resources

We have presented key resources for many eukaryotic organisms and their genome-sequencing websites. An excellent starting point is the Ensembl website (▸http://www.ensembl.org/), which currently includes gateways for the mouse, rat, zebrafish, fugu, mosquito, and other genomes.

Discussion Questions

[16-1] If there were no repetitive DNA of any kind, how would the genomes of various eukaryotes (human, mouse, a plant, a parasite) compare in terms of size, gene content, gene order, nucleotide composition, or other features?

[16-2] If someone gave you 1 Mb of genomic DNA sequence from a eukaryote, how could you identify the species? (Assume you cannot use BLAST to directly identify the species.) What features distinguish the genomic DNA sequence of a protozoan parasite from an insect or a fish?

Problems

[16-1] (a) Retrieve a typical *Arabadopsis thaliana* bacterial artificial chromosome (BAC) from Entrez. (E.g., choose BAC T18A20, GenBank accession AC009324.)

- Note the approximate size (in kilobases). Is this a large or a small BAC?
- Note the approximate number of protein products in it. Bacteria have about one gene per kilobase. How many genes are there per kilobase in this eukaryotic DNA?

(b) Go to the ORF Finder at NCBI:

- From the main page, look at the left sidebar. Choose "Tools for data mining"; then you will see the ORF Finder.
- Alternatively, from the main page, look at the left sidebar at the top. Choose "Site map" and you will also find a link to the ORF Finder.
- Paste in the accession number for your BAC. Click OrfFind.

(c) At the ORF Finder at NCBI, Click on the largest ORF.

- How many amino acids long is it?
- What is its molecular weight (in kilodaltons)?
- Is this protein small, average, or large?
- From which strand of the BAC is this putative gene transcribed? Overall, are there more ORFs on the top or bottom strand or is it about the same?

(d) Using the ORF Finder at NCBI, BLAST search the ORF of (c) using the default parameters that are given to you.

- Note that the results page is NOT updated automatically so you may need to reload your page.
- This BLAST result reveals many matches to *Arabidopsis* proteins. However, note that if you do a standard blastp search using this ORF as a query, you will find matches to many dozens of species. Also you will see a match to the Conserved Domain Database. Thus, the BLAST tool within OrfFinder is not as thorough as a regular BLAST search.

[16-2] Human centromeres typically contain several thousand base pairs of a 171-bp repeat called α-satellite (accession X07685). First perform a blastn search against the nonredundant database. What kinds of database matches do you observe? Second, restrict your BLAST search to nonhumans. (In the options section that allows you to limit by Entrez query, try typing "satellite NOT human[organism]."). Are there matches in primates, rodents, or plants? Why might centromeric repeats have this phylogenetic distribution; would you expect each species to have its own, unique centromeric signature?

[16-3] A universal minicircle binding protein (GenBank accession A54598) has been purified from a trypanosome that infects insects, *Crithidia fasciculata*. A blastp search reveals that there are homologous proteins in plants, fungi, and metazoans (such as the worm *Caenorhabditis elegans*). How is this protein named in various organisms? What is its presumed function? What is its domain called in the Conserved Domain Database?

[16-4] *Leishmania major* has repetitive DNA elements (e.g., accession AF421497). How can you decide how common this element is and where it is localized (e.g., to a particular chromosome or to a chromosomal region).

SELF-TEST QUIZ

[16-1] The C value paradox is that
(a) the nucleotide C is underrepresented in some genomes
(b) the genome size of various eukaryotes correlates poorly with the number of protein-coding genes of the organism
(c) the genome size of various eukaryotes correlates poorly with the biological complexity of the organism
(d) the genome size of various eukaryotes correlates poorly with the evolutionary age of the organism

[16-2] Hundreds or thousands of sequence repeats, each consisting of a unit of about 4 to 8 nucleotides, are commonly found where?
(a) in interspersed repeats
(b) in processed pseudogenes
(c) in telomeres
(d) in segmentally duplicated regions

[16-3] You are sequencing the genome of a newly described organism (a slime mold). What is likely to happen if you use RepeatMasker to assess its repetitive DNA content? You set the default setting of RepeatMasker to the settings for human DNA.
(a) RepeatMasker should successfully identify essentially all of the repetitive DNA. Various repetitive DNA elements are similar enough between organisms to allow this software to work on your slime mold DNA.
(b) RepeatMasker should identify most of the repetitive DNA. However, because some types of repeats are species-specific, it is likely that there will be many false positive and false negative results.
(c) RepeatMasker would fail to identify most of the repetitive DNA. Most types of repeats are highly species-specific. It is necessary for you to train the RepeatMasker algorithm on your slime mold DNA in order for the program to work.
(d) It is not possible to predict, because repetitive DNA may or may not be variable between organisms.

[16-4] It is extremely difficult for intrinsic (ab initio) gene-finding algorithms to predict protein-coding genes in eukaryotic genomic DNA . What is the main problem?
(a) exon/intron borders are hard to predict.
(b) introns may be many kilobases in length.
(c) the GC content of coding regions is not always differentiated from the GC content of noncoding regions.
(d) all of the above.

[16-5] The genomes of two distinct eukaryotic species can sometimes merge to create an entirely new species.
(a) True
(b) False

[16-6] The *Giardia lamblia* genome is unusual because
(a) it contains hardly any transposable elements or introns
(b) it is circular
(c) it contains extremely little non-repetitive DNA
(d) its AT content is nearly 80%

[16-7] Plants have been among the most successful organisms on earth, diversifying into thousands of species. The rice and *Arabidopsis* genomes were the first plant genomes to be sequenced. Comparing them to metazoan genomes in broad terms, these plant genomes
(a) are basically comparable in terms of GC content, range of genome sizes, and range of number of genes
(b) are distinguished because the plant genomes have less repetitive DNA

(c) are probably distinguished because the plant genomes consistently underwent a more rapid rate of evolution
(d) are fundamentally distinguished because the plant genomes incorporate atypical nucleotides

[16-8] How are the worm (*C. elegans*) and yeast (*S.cerevisiae*) genomes different?
(a) The worm genome includes significantly more genes encoding multidomain proteins
(b) The core set of worm proteins is about ten times larger than the core set of yeast proteins
(c) The GC content of the worm genome is considerably higher.
(d) The fraction of genes in the worm genome that encode proteins without known orthologs is considerably higher.

[16-9] How are the mouse and human genomes different?
(a) the mouse genome has a lower GC content
(b) the mouse genome has more protein-coding genes
(c) the mouse genome has undergone specific expansions of genes encoding particular protein families such as olfactory receptors
(d) the mouse genome has fewer telomeric repeats per chromosome, on average

Suggested Reading

The classic studies of Britten and Kohne (1968) are highly recommended for explanations of repetitive DNA. For eukaryotic gene prediction, a review of software programs by Makarov (2002) is recommended.

David Searls (2000) reviews bioinformatics tools for whole genes in a chapter that discusses a broad variety of subjects (e.g., BLAST and ESTs).

References

Adam, R. D. Biology of *Giardia lamblia*. *Clin. Microbiol. Rev.* **14,** 447–475 (2001).

Adams, M. D., et al. The genome sequence of *Drosophila melanogaster*. *Science* **287,** 2185–2195 (2000).

Akman, L., et al. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nat. Genet.* **32,** 402–407 (2002).

Allen, K. D. Assaying gene content in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **99,** 9568–9572 (2002).

Ambros, V. microRNAs: Tiny regulators with great potential. *Cell* **107,** 823–826 (2001).

Amor, D. J., and Choo, K. H. Neocentromeres: Role in human disease, evolution, and centromere study. *Am. J. Hum. Genet.* **71,** 695–714 (2002).

The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).

Antonarakis, S. E., et al. Factor VIII gene inversions in severe hemophilia A: Results of an international consortium study. *Blood* **86,** 2206–2212 (1995).

Aparicio, S., et al. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297,** 1301–1310 (2002).

Arkhipova, I., and Meselson, M. Transposable elements in sexual and ancient asexual taxa. *Proc. Natl. Acad. Sci. USA* **97,** 14473–14477 (2000).

Arkhipova, I. R., and Morrison, H. G. Three retrotransposon families in the genome of *Giardia lamblia:* Two telomeric, one dead. *Proc. Natl. Acad. Sci. USA* **98,** 14497–14502 (2001).

Avramova, Z. V. Heterochromatin in animals and plants. Similarities and differences. *Plant Physiol.* **129,** 40–49 (2002).

Azzalin, C. M., Nergadze, S. G., and Giulotto, E. Human intrachromosomal telomeric-like repeats: Sequence organization and mechanisms of origin. *Chromosoma* **110,** 75–82 (2001).

Bahl, A., et al. PlasmoDB: The *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Res.* **31,** 212–215 (2003).

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., and Eichler, E. E. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11,** 1005–1017 (2001).

Baldauf, S. L., Roger, A. J., Wenk-Siefert, I., and Doolittle, W. F. A kingdom-level phylogeny of eukaryotes based on combined protein data. *Science* **290,** 972–977 (2000).

Bennetzen, J. L., and Freeling, M. The unified grass genome: Synergy in synteny. *Genome Res.* **7,** 301–306 (1997).

Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).

Ben-Yehuda, S., Rudner, D. Z., and Losick, R. RacA, a bacterial protein that anchors chromosomes to the cell poles. *Science* **19,** 19 (2002).

Betrán, E., and Long, M. Expansion of genome coding regions by acquisition of new genes. *Genetica* **115,** 65–80 (2002).

Blackburn, E. H., et al. Recognition and elongation of telomeres by telomerase. *Genome* **31,** 553–560 (1989).

Blake, J. A., Richardson, J. E., Bult, C. J., Kadin, J. A., and Eppig, J. T. MGD: The Mouse Genome Database. *Nucleic Acids Res.* **31,** 193–195 (2003).

Blattner, F. R., et al. The complete genome sequence of *Escherichia coli* K-12. *Science* **277,** 1453–1474 (1997).

Blaxter, M. *Caenorhabditis elegans* is a nematode. *Science* **282,** 2041–2046 (1998).

Blaxter, M. Molecular systematics: Counting angels with DNA. *Nature* **421,** 122–124 (2003).

Bray, N., Dubchak, I., and Pachter, L. AVID: A Global Alignment Program. *Genome Res.* **13,** 97–102 (2003).

Britten, R. J., and Kohne, D. E. Repeated sequences in DNA. *Science* **161,** 529–540 (1968).

Brooks, D. R., and Isaac, R. E. Functional genomics of parasitic worms: The dawn of a new era. *Parasitol. Int.* **51,** 319–325 (2002).

Budiansky, S. Creatures of our own making. *Science* **298,** 80–86 (2002).

Buell, C. R. Current status of the sequence of the rice genome and prospects for finishing the first monocot genome. *Plant Physiol.* **130,** 1585–1586 (2002).

Burge, C., and Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).

Burge, C. B., and Karlin, S. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8,** 346–354 (1998).

Cameron, R. A., et al. A sea urchin genome project: Sequence scan, virtual map, and additional resources. *Proc. Natl. Acad. Sci. USA* **97,** 9514–9518 (2000).

Carlton, J. M., et al. Profiling the malaria genome: A gene survey of three species of malaria parasite with comparison to other apicomplexan species. *Mol. Biochem. Parasitol.* **118,** 201–210 (2001).

Carlton, J. M., et al. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii. Nature* **419,** 512–519 (2002).

Cavalier-Smith, T. A revised six-kingdom system of life. *Biol. Rev. Camb. Philos. Soc.* **73,** 203–266 (1998).

Cavalier-Smith, T. Origins of the machinery of recombination and sex. *Heredity* **88,** 125–141 (2002).

Celniker, S. E., et al. Finishing a whole-genome shotgun: Release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol.* **3,** RESEARCH0079-9 (2002).

Centers for Disease Control and Prevention. Provisional surveillance summary of the West Nile virus epidemic–United States, January–November 2002. *MMWR* **51,** 1129–1133 (2002).

Chan, P., Simon-Chazottes, D., Mattei, M. G., Guenet, J. L., and Salier, J. P. Comparative mapping of lipocalin genes in human and mouse: The four genes for complement C8 gamma chain, prostaglandin-D-synthase, oncogene-24p3, and progestagen-associated endometrial protein map to HSA9 and MMU2. *Genomics* **23,** 145–150 (1994).

Choo, K. H. Domain organization at the centromere and neocentromere. *Dev. Cell.* **1,** 165–177 (2001).

Claverie, J. M. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6,** 1735–1744 (1997).

Claverie, J. M. Gene number. What if there are only 30,000 human genes? *Science* **291,** 1255–1257 (2001).

Comai, L. Genetic and epigenetic interactions in allopolyploid plants. *Plant Mol. Biol.* **43,** 387–399 (2000).

Cowman, A. F., and Crabb, B. S. The *Plasmodium falciparum* genome–a blueprint for erythrocyte invasion. *Science* **298,** 126–128 (2002).

Cox, F. E. History of human parasitology. *Clin. Microbiol. Rev.* **15,** 595–612 (2002).

Crowe, M. L., Serizet, C., Thareau, V., Aubourg, S., Rouze, P., Hilson, P., Beynon, J., Weisbeek, P., van Hummelen, P., Reymond, P., Paz-Ares, J., Nietfeld, W., Trick, M. CATMA: a complete *Arabidopsis* GST database. *Nucleic Acids Res.* **31,** 156–158 (2003).

Cummings, C. J., and Zoghbi, H. Y. Trinucleotide repeats: Mechanisms and pathophysiology. *Annu. Rev. Genomics Hum. Genet.* **1,** 281–328 (2000).

Cuvier, G. *The Animal Kingdom, Arranged According to its Organization.* William S. Orr & Co., London, 1849.

Dehal, P., et al. Human chromosome 19 and related regions in mouse: Conservative and lineage-specific evolution. *Science* **293,** 104–111 (2001).

Dehal, P., et al. The draft genome of *Ciona intestinalis:* Insights into chordate and vertebrate origins. *Science* **298,** 2157–2167 (2002).

Dermitzakis, E. T., et al. Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature* **420,** 578–582 (2002).

Dessen, P., et al. Paramecium genome survey: A pilot project. *Trends Genet.* **17,** 306–308 (2001).

Dewald, G., et al. The human complement C8G gene, a member of the lipocalin gene family: Polymorphisms and mapping to chromosome 9q34.3. *Ann. Hum. Genet.* **60,** 281–291 (1996).

Donelson, J. E. Genome research and evolution in trypanosomes. *Curr. Opin. Genet. Dev.* **6,** 699–703 (1996).

Dong, Q., Roy, L., Freeling, M., Walbot, V., and Brendel, V. ZmDB, an integrated database for maize genome research. *Nucleic Acids Res.* **31,** 244–247 (2003).

Douglas, S. E., and Penny, S. L. The plastid genome of the cryptophyte alga, *Guillardia theta:* Complete sequence and conserved synteny groups confirm its common ancestry with red algae. *J. Mol. Evol.* **48,** 236–244 (1999).

Douglas, S., et al. The highly reduced genome of an enslaved algal nucleus. *Nature* **410,** 1091–1096 (2001).

Echols, N., et al. Comprehensive analysis of amino acid and nucleotide composition in eukaryotic genomes, comparing genes and pseudogenes. *Nucleic Acids Res.* **30,** 2515–2523 (2002).

Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* **2,** 919–929 (2001).

Eddy, S. R. Computational genomics of noncoding RNA genes. *Cell* **109,** 137–140 (2002).

El-Sayed, N. M., Hegde, P., Quackenbush, J., Melville, S. E., and Donelson, J. E. The African trypanosome genome. *Int. J. Parasitol.* **30,** 329–345 (2000).

Embley, T. M., and Hirt, R. P. Early branching eukaryotes? *Curr. Opin. Genet. Dev.* **8,** 624–629 (1998).

Fan, Y., Linardopoulou, E., Friedman, C., Williams, E., and Trask, B. J. Genomic structure and evolution of the ancestral chromosome fusion site in 2q13–2q14.1 and paralogous regions on other human chromosomes. *Genome Res.* **12,** 1651–1662 (2002).

Feng, Q., et al. Sequence and analysis of rice chromosome 4. *Nature* **420,** 316–320 (2002).

Fields, S., Kohara, Y., and Lockhart, D. J. Functional genomics. *Proc. Natl. Acad. Sci. USA* **96,** 8825–8256 (1999).

Florens, L., et al. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature* **419,** 520–526 (2002).

The FlyBase Consortium. The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.* **31,** 172–175 (2003).

Fraser, C. M., Eisen, J. A., and Salzberg, S. L. Microbial genome sequencing. *Nature* **406,** 799–803 (2000).

Frazer, K. A., Elnitski, L., Church, D. M., Dubchak, I., and Hardison, R. C. Cross-species sequence comparisons: A review of methods and available resources. *Genome Res.* **13,** 1–12 (2003).

Gardner, M. J., et al. Genome sequence of the human malaria parasite *Plasmodium falciparum. Nature* **419,** 498–511 (2002).

Gelbart, W. M. Databases in genomic research. *Science* **282,** 659–661 (1998).

The *C. elegans* Sequencing Consortium. Genome sequence of the nematode *C. elegans:* A platform for investigating biology. *Science* **282,** 2012–2018 (1998).

Gentles, A. J., and Karlin, S. Genome-scale compositional comparisons in eukaryotes. *Genome Res.* **11,** 540–546 (2001).

Gilson, P. R., and McFadden, G. I. A grin without a cat. *Nature* **410,** 1040–1041 (2001).

Gilson, P. R., and McFadden, G. I. Jam packed genomes–a preliminary, comparative analysis of nucleomorphs. *Genetica* **115,** 13–28 (2002).

Glockner, G., et al. Sequence and analysis of chromosome 2 of *Dictyostelium discoideum. Nature* **418,** 79–85 (2002).

Goff, S. A., et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296,** 92–100 (2002).

Graur, D., and Li, W.-H. *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, MA, 2000.

Gregory, S. G., et al. A physical map of the mouse genome. *Nature* **418,** 743–750 (2002).

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S. R. Rfam: An RNA family database. *Nucleic Acids Res.* **31,** 439–441 (2003).

Gutierrez, G., Ganfornina, M. D., and Sanchez, D. Evolution of the lipocalin family as inferred from a protein sequence phylogeny. *Biochim. Biophys. Acta* **1482,** 35–45 (2000).

Hall, A. E., Fiebig, A., and Preuss, D. Beyond the *Arabidopsis* genome: Opportunities for comparative genomics. *Plant Physiol.* **129,** 1439–1447 (2002).

Hancock, J. M. Genome size and the accumulation of simple sequence repeats: Implications of new data from genome sequencing projects. *Genetica* **115,** 93–103 (2002).

Hardison, R. C., Oeltjen, J., and Miller, W. Long human–mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7,** 959–966 (1997).

Harris, T. W., et al. WormBase: A cross-species database for comparative genomics. *Nucleic Acids Res.* **31,** 133–137 (2003).

Harrison, P. M., and Gerstein, M. Studying genomes through the aeons: Protein families, pseudogenes and proteome evolution. *J. Mol. Biol.* **318,** 1155–1174 (2002).

Harrison, P. M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. A question of size: The eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.* **30,** 1083–1090 (2002).

Hartl, D. L. Molecular melodies in high and low C. *Nat. Rev. Genet.* **1,** 145–149 (2000).

Hebert, P. D. N., Cywinska, A., Ball, S. L., and deWaard, J. R. Biological identification through DNA barcodes. *Proc. R. Soc. Lond.* **270,** 313–321 (2003).

Hedges, S. B., and Kumar, S. Genomics. Vertebrate genomes compared. *Science* **297,** 1283–1285 (2002).

Henrich, T., et al. MEPD: A Medaka gene expression pattern database. *Nucleic Acids Res.* **31,** 72–74 (2003).

Hill, A. A., Hunter, C. P., Tsung, B. T., Tucker-Kellogg, G., and Brown, E. L. Genomic analysis of gene expression in *C. elegans. Science* **290,** 809–812 (2000).

Hoffman, S. L., Subramanian, G. M., Collins, F. H., and Venter, J. C. *Plasmodium,* human and *Anopheles* genomics and malaria. *Nature* **415,** 702–709 (2002).

Holland, P. W. Ciona. *Curr. Biol.* **12,** R609 (2002).

Holt, R. A., et al. The genome sequence of the malaria mosquito *Anopheles gambiae. Science* **298,** 129–149 (2002).

Hou, G., Le Blancq, S. M., Yaping, E., Zhu, H., and Lee, M. G. Structure of a frequently rearranged rRNA-encoding

chromosome in *Giardia lamblia*. *Nucleic Acids Res.* **23,** 3310–3317 (1995).

IJdo, J. W., Baldini, A., Ward, D. C., Reeders, S. T., and Wells, R. A. Origin of human chromosome 2: An ancestral telomere-telomere fusion. *Proc. Natl. Acad. Sci. USA* **88,** 9051–9055 (1991a).

IJdo, J. W., Wells, R. A., Baldini, A., and Reeders, S. T. Improved telomere detection using a telomere repeat probe (TTAGGG)n generated by PCR. *Nucleic Acids Res.* **19,** 4780 (1991b).

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

Jarstfer, M. B., and Cech, T. R. Effects of nucleotide analogues on *Euplotes aediculatus* telomerase processivity: Evidence for product-assisted translocation. *Biochemistry* **41,** 151–161 (2002).

Jentsch, S., Tobler, H., and Muller, F. New telomere formation during the process of chromatin diminution in *Ascaris suum*. *Int. J. Dev. Biol.* **46,** 143–148 (2002).

Johnson, P. J. Spliceosomal introns in a deep-branching eukaryote: The splice of life. *Proc. Natl. Acad. Sci. USA* **99,** 3359–3361 (2002).

Jomaa, H., et al. Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. *Science* **285,** 1573–1576 (1999).

Jones, K. L. *Smith's Recognizable Patterns of Human Malformation,* W. B. Saunders, New York, 1997.

Jurka, J. Repeats in genomic DNA: Mining and meaning. *Curr. Opin. Struct. Biol.* **8,** 333–337 (1998).

Jurka, J. Repbase update: A database and an electronic journal of repetitive elements. *Trends Genet.* **16,** 418–420 (2000).

Kamath, R. S., et al. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421,** 231–237 (2003).

Kapranov, P., et al. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296,** 916–919 (2002).

Karlowski, W. M., Schoof, H., Janakiraman, V., Stuempflen, V., and Mayer, K. F. MOsDB: An integrated information resource for rice genomics. *Nucleic Acids Res.* **31,** 190–192 (2003).

Kidwell, M. G. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115,** 49–63 (2002).

Kim, S. K. Functional genomics: The worm scores a knockout. *Curr. Biol.* **11,** R85–R87 (2001).

Kissinger, J. C., Gajria, B., Li, L., Paulsen, I. T., and Roos, D. S. ToxoDB: Accessing the *Toxoplasma gondii* genome. *Nucleic Acids Res.* **31,** 234–236 (2003).

Knight, J. All genomes great and small. *Nature* **417,** 374–376 (2002).

Knight, R. D., Freeland, S. J., and Landweber, L. F. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. *Genome Biol.* **2** (2001).

Kotzot, D. Complex and segmental uniparental disomy (UPD): Review and lessons from rare chromosomal complements. *J. Med. Genet.* **38,** 497–507 (2001).

Ku, H. M., Vision, T., Liu, J., and Tanksley, S. D. Comparing sequenced segments of the tomato and *Arabidopsis* genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. *Proc. Natl. Acad. Sci. USA* **97,** 9121–9126 (2000).

Kumar, S., and Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* **392,** 917–920 (1998).

Lacazette, E., Gachon, A. M., and Pitiot, G. A novel human odorant-binding protein gene family resulting from genomic duplicons at 9q34: Differential expression in the oral and genital spheres. *Hum. Mol. Genet.* **9,** 289–301 (2000).

Lamblin, A. F., et al. MtDB: A database for personalized data mining of the model legume *Medicago truncatula* transcriptome. *Nucleic Acids Res.* **31,** 196–201 (2003).

Lasonder, E., et al. Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry. *Nature* **419,** 537–542 (2002).

Lee, H. S., and Chen, Z. J. Protein-coding genes are epigenetically regulated in *Arabidopsis* polyploids. *Proc. Natl. Acad. Sci. USA* **98,** 6753–6758 (2001).

Lewis, M. S., and Pikaard, C. S. Restricted chromosomal silencing in nucleolar dominance. *Proc. Natl. Acad. Sci. USA* **98,** 14536–14540 (2001).

Lloyd, D., and Harris, J. C. *Giardia:* Highly evolved parasite or early branching eukaryote? *Trends Microbiol.* **10,** 122–127 (2002).

Long, C. A., and Hoffman, S. L. Malaria—from infants to genomics to vaccines. *Science* **297,** 345–347 (2002).

Lorite, P., Carrillo, J. A., and Palomeque, T. Conservation of (TTAGG)(n) telomeric sequences among ants (Hymenoptera, Formicidae). *J. Hered.* **93,** 282–285 (2002).

Lowe, T. M., and Eddy, S. R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).

Makalowski, W. Genomic scrap yard: How genomes utilize all that junk. *Gene* **259,** 61–67 (2000).

Makarov, V. Computer programs for eukaryotic gene prediction. *Brief. Bioinform.* **3,** 195–199 (2002).

Margulis, L., and Schwartz, K. V. *Five Kingdoms. An Illustrated Guide to the Phyla of Life on Earth.* W. H. Freeman, New York, 1998.

Martin, C. L., et al. The evolutionary origin of human subtelomeric homologies—or where the ends begin. *Am. J. Hum. Genet.* **70,** 972–984 (2002).

Matthews, D. E., Carollo, V. L., Lazo, G. R., and Anderson, O. D. GrainGenes, the genome database for small-grain crops. *Nucleic Acids Res.* **31,** 183–186 (2003).

Mayor, C., et al. VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16,** 1046–1047 (2000).

McArthur, A. G., et al. The *Giardia* genome project database. *FEMS Microbiol. Lett.* **189,** 271–273 (2000).

McConkey, E. H., and Varki, A. A primate genome project deserves high priority. *Science* **289,** 1295–1296 (2000).

McCormick-Graham, M., and Romero, D. P. A single telomerase RNA is sufficient for the synthesis of variable telomeric DNA repeats in ciliates of the genus *Paramecium*. *Mol. Cell. Biol.* **16,** 1871–1879 (1996).

McKnight, T. D., Fitzgerald, M. S., and Shippen, D. E. Plant telomeres and telomerases. A review. *Biochemistry (Mosc.)* **62,** 1224–1231 (1997).

Melek, M., Davis, B. T., and Shippen, D. E. Oligonucleotides complementary to the *Oxytricha nova* telomerase RNA delineate the template domain and uncover a novel mode of primer utilization. *Mol. Cell. Biol.* **14,** 7827–7838 (1994).

Meyerowitz, E. M. Plants compared to animals: The broadest comparative study of development. *Science* **295,** 1482–1485 (2002).

Misra, S., et al. Annotation of the *Drosophila melanogaster* euchromatic genome: A systematic review. *Genome Biol.* **3,** RESEARCH0083-3 (2002).

Moller-Jensen, J., Jensen, R. B., Lowe, J., and Gerdes, K. Prokaryotic DNA segregation by an actin-like filament. *EMBO J.* **21,** 3119–3127 (2002).

Morgenstern, B., et al. Exon discovery by genomic sequence alignment. *Bioinformatics* **18,** 777–787 (2002).

Morris, J. C., et al. Replication of kinetoplast DNA: An update for the new millennium. *Int. J. Parasitol.* **31,** 453–458 (2001).

Müller, F., and Tobler, H. Chromatin diminution in the parasitic nematodes *Ascaris suum* and *Parascaris univalens*. *Int. J. Parasitol.* **30,** 391–399 (2000).

Mural, R. J. Current status of computational gene finding: A perspective. *Methods Enzymol.* **303,** 77–83 (1999).

Mural, R. J., et al. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science* **296,** 1661–1671 (2002).

Murthy, V. L., and Rose, G. D. RNABase: An annotated database of RNA structures. *Nucleic Acids Res.* **31,** 502–504 (2003).

Myler, P. J., et al. *Leishmania major* Friedlin chromosome 1 has an unusual distribution of protein-coding genes. *Proc. Natl. Acad. Sci. USA* **96,** 2902–2906 (1999).

Myler, P. J., et al. Genomic organization and gene function in *Leishmania*. *Biochem. Soc. Trans.* **28,** 527–531 (2000).

Nadeau, J. H., and Taylor, B. A. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. USA* **81,** 814–818 (1984).

Nanda, I., et al. Distribution of telomeric (TTAGGG)(n) sequences in avian chromosomes. *Chromosoma* **111,** 215–227 (2002).

Nixon, J. E., et al. A spliceosomal intron in *Giardia lamblia*. *Proc. Natl. Acad. Sci. USA* **99,** 3701–3705 (2002).

Novichkov, P. S., Gelfand, M. S., and Mironov, A. A. Gene recognition in eukaryotic DNA by comparison of genomic sequences. *Bioinformatics* **17,** 1011–1018 (2001).

Ostertag, E. M., and Kazazian, H. H., Jr. Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* **11,** 2059–2065 (2001).

Pardue, M. L., DeBaryshe, P. G., and Lowenhaupt, K. Another protozoan contributes to understanding telomeres and transposable elements. *Proc. Natl. Acad. Sci. USA* **98,** 14195–14197 (2001).

Passarge, E., Horsthemke, B., and Farber, R. A. Incorrect use of the term synteny. *Nat. Genet.* **23,** 387 (1999).

Pelosi, P. Perireceptor events in olfaction. *J. Neurobiol.* **30,** 3–19 (1996).

Pevsner, J., Hou, V., Snowman, A. M., and Snyder, S. H. Odorant-binding protein. Characterization of ligand binding. *J. Biol. Chem.* **265,** 6118–6125 (1990).

Philippe, H., and Laurent, J. How good are deep phylogenetic trees? *Curr. Opin. Genet. Dev.* **8,** 616–623 (1998).

Redi, C. A., Garagna, S., Zacharias, H., Zuccotti, M., and Capanna, E. The other chromatin. *Chromosoma* **110,** 136–147 (2001).

Reese, M. G., Kulp, D., Tammana, H., and Haussler, D. Genie–gene finding in *Drosophila melanogaster*. *Genome Res.* **10,** 529–538 (2000).

Reich, D. E., Lander, E. S., Waterson, R., Pääbo, S., Ruvolo, M., and Varki, A. Sequencing the chimpanzee genome. On-line white paper (not dated), ▸www.hgsc.bcm.tmc.edu/projects/chimpanzee/ChimpGenome2.pdf.

Reyes, A., Pesole, G., and Saccone, C. Long-branch attraction phenomenon and the impact of among-site rate variation on rodent phylogeny. *Gene* **259,** 177–187 (2000).

Rhee, S. Y., et al. The Arabidopsis Information Resource (TAIR): A model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* **31,** 224–228 (2003).

Rogic, S., Mackworth, A. K., and Ouellette, F. B. Evaluation of gene-finding programs on mammalian sequences. *Genome Res.* **11,** 817–832 (2001).

Rubin, G. M., and Lewis, E. B. A brief history of *Drosophila*'s contributions to genome research. *Science* **287,** 2216–2218 (2000).

Rubin, G. M., et al. Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).

Rudd, S., Mewes, H. W., and Mayer, K. F. Sputnik: A database platform for comparative plant genomics. *Nucleic Acids Res.* **31,** 128–132 (2003).

Ruvkun, G. Molecular biology. Glimpses of a tiny RNA world. *Science* **294,** 797–799 (2001).

Ryan, J. F., and Finnerty, J. R. CnidBase: The Cnidarian Evolutionary Genomics Database. *Nucleic Acids Res.* **31,** 159–163 (2003).

Salier, J. P. Chromosomal location, exon/intron organization and evolution of lipocalin genes. *Biochim. Biophys. Acta* **1482,** 25–34 (2000).

Samson, D., et al. GenoPlante-Info (GPI): A collection of databases and bioinformatics resources for plant genomics. *Nucleic Acids Res.* **31,** 179–182 (2003).

Sargis, E. J. Paleontology. Primate origins nailed. *Science* **298,** 1564–1565 (2002).

Sasaki, T., et al. The genome sequence and structure of rice chromosome 1. *Nature* **420,** 312–316 (2002).

Schlötterer, C., and Harr, B. *Drosophila virilis* has long and highly polymorphic microsatellites. *Mol. Biol. Evol.* **17,** 1641–1646 (2000).

Schwartz, S., et al. PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10,** 577–586 (2000).

Searls, D. B. *Bioinformatics Tools for Whole Genomes.* Annual Reviews, Palo Alto, CA, 2000, pp. 251–279.

Seo, H. C., et al. Miniature genome in the marine chordate *Oikopleura dioica. Science* **294,** 2506 (2001).

Shapiro, T. A., and Englund, P. T. The structure and replication of kinetoplast DNA. *Annu. Rev. Microbiol.* **49,** 117–143 (1995).

Shippen-Lentz, D., and Blackburn, E. H. Telomere terminal transferase activity from *Euplotes crassus* adds large numbers of TTTTGGGG repeats onto telomeric primers. *Mol. Cell. Biol.* **9,** 2761–2764 (1989).

Simillion, C., Vandepoele, K., Van Montagu, M. C., Zabeau, M., and Van de Peer, Y. The hidden duplication past of *Arabidopsis thaliana. Proc. Natl. Acad. Sci. USA* **99,** 13627–13632 (2002).

Simpson, A. G., MacQuarrie, E. K., and Roger, A. J. Eukaryotic evolution: Early origin of canonical introns. *Nature* **419,** 270 (2002).

Slijepcevic, P. Telomeres and mechanisms of Robertsonian fusion. *Chromosoma* **107,** 136–140 (1998a).

Slijepcevic, P. Telomere length regulation—a view from the individual chromosome perspective. *Exp. Cell Res.* **244,** 268–274 (1998b).

Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9,** 657–663 (1999).

Snel, B., Bork, P., and Huynen, M. A. Genome phylogeny based on gene content. *Nat. Genet.* **21,** 108–110 (1999).

Sonnhammer, E. L., and Durbin, R. Analysis of protein domain families in *Caenorhabditis elegans. Genomics* **46,** 200–216 (1997).

Sprague, J., et al. The Zebrafish Information Network (ZFIN): The zebrafish model organism database. *Nucleic Acids Res.* **31,** 241–243 (2003).

Stauffer, R. L., Walker, A., Ryder, O. A., Lyons-Weiler, M., and Hedges, S. B. Human and ape molecular clocks and constraints on paleontological hypotheses. *J. Hered.* **92,** 469–474 (2001).

Stechmann, A., and Cavalier-Smith, T. Rooting the eukaryote tree by using a derived gene fusion. *Science* **297,** 89–91 (2002).

Stein, L. Genome annotation: From sequence to biology. *Nat. Rev. Genet.* **2,** 493–503 (2001).

Szymanski, M., Erdmann, V. A., and Barciszewski, J. Noncoding regulatory RNAs database. *Nucleic Acids Res.* **31,** 429–431 (2003).

Tzafrir, I., et al. The *Arabidopsis* SeedGenes Project. *Nucleic Acids Res.* **31,** 90–93 (2003).

Upcroft, P., Chen, N., and Upcroft, J. A. Telomeric organization of a variable and inducible toxin gene family in the ancient eukaryote *Giardia duodenalis. Genome Res* **7,** 37–46 (1997).

Van de Peer, Y., Baldauf, S. L., Doolittle, W. F., and Meyer, A. An updated and comprehensive rRNA phylogeny of (crown) eukaryotes based on rate-calibrated evolutionary distances. *J. Mol. Evol.* **51,** 565–576 (2000).

Vellai, T., and Vida, G. The origin of eukaryotes: The difference between prokaryotic and eukaryotic cells. *Proc. R. Soc. Lond. B Biol. Sci.* **266,** 1571–1577 (1999).

Walbot, V. *Arabidopsis thaliana* genome. A green chapter in the book of life. *Nature* **408,** 794–795 (2000).

Wang, D. Y., Kumar, S., and Hedges, S. B. Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc. R. Soc. Lond. B Biol. Sci.* **266,** 163–171 (1999).

Waterston, R. H., et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).

Watt, W. B., and Dean, A. M. Molecular-functional studies of adaptive genetic variation in prokaryotes and eukaryotes. *Annu. Rev. Genet.* **34,** 593–622 (2000).

Williams, B. A., Hirt, R. P., Lucocq, J. M., and Embley, T. M. A mitochondrial remnant in the microsporidian *Trachipleistophora hominis. Nature* **418,** 865–869 (2002).

Wirth, D. F. Biological revelations. *Nature* **419,** 495–496 (2002).

Wolfe, K. H., Morden, C. W., and Palmer, J. D. Function and evolution of a minimal plastid genome from a nonphotosynthetic parasitic plant. *Proc. Natl. Acad. Sci. USA* **89,** 10648–10652 (1992).

Wu, J., et al. A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell* **14,** 525–535 (2002).

Xuan, Z., McCombie, W. R., and Zhang, M. Q. GFScan: A gene family search tool at genomic DNA level. *Genome Res.* **12,** 1142–1149 (2002).

Xuan, Z., Wang, J., and Zhang, M. Q. Computational comparison of two mouse draft genomes and the human golden path. *Genome Biol.* **4** (2003).

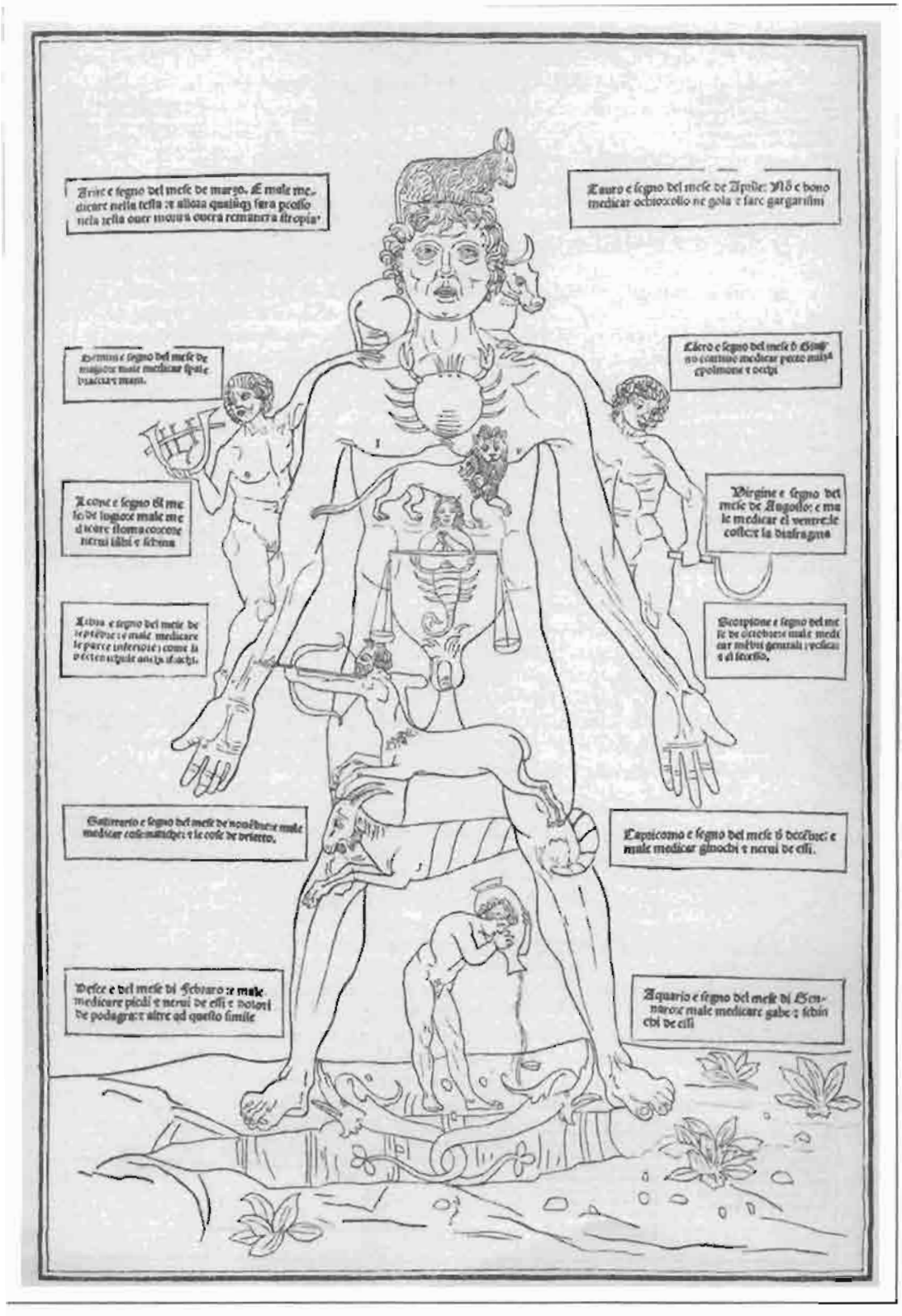
Yu, J., et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296,** 79–92 (2002).

Yuan, Q., et al. The TIGR rice genome annotation resource: Annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res.* **31,** 229–233 (2003).

Zdobnov, E. M., et al. Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298,** 149–159 (2002).

Zuegge, J., Ralph, S., Schmuker, M., McFadden, G. I., and Schneider, G. Deciphering apicoplast targeting signals–feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene* **280,** 19–26 (2001).

This Page Intentionally Left Blank



The "zodiac man" from Johannes de Ketham's Fasciculus medicinae *(1493; see Singer, 1925) shows the astrological signs superimposed on the human body. For many centuries this kind of figure was used as a guide to phlebotomy (blood-letting). The organs of the body are assigned celestial signs, and when the moon was in the sign of the patient's organ the physician was to avoid treatment. From the top left, the first labels read: "Aries is the sign of March; it is bad to treat the head, and who is then struck on the head either dies or remains injured." "Gemini is the sign of May; it is bad to treat or to phlebotomize the shoulders, arms and hands."*

17

Human Genome

INTRODUCTION

The human genome is the complete set of DNA in *Homo sapiens*. This DNA encodes the proteins that define who we are as biological entities. Through the genomic DNA, protein-coding genes are expressed that form the architecture of the trillions of cells that comprise each of our bodies. It is variations in the genome that account for the differences between people, from physical features to personality to disease states.

The sequencing of the human genome is a triumph of science. It follows almost 50 years exactly after the publication of the double-stranded helical structure of DNA by Crick and Watson (1953). The genome sequence has been achieved through an international collaboration involving hundreds of scientists. (In the case of the publicly funded version, this was the International Human Genome Sequencing Consortium (IHGSC), described below.) This project could not have been possible without fundamental advances in the emerging fields of bioinformatics and genomics.

This chapter is divided in three parts. First, we summarize some of the major findings of the human genome project. Second, we will introduce web-based resources for the study of the human genome at three sites: the National Center for Biotechnology Information (NCBI), the Ensembl project, and the genome center at the University of California, Santa Cruz.

On February 15, 2001 the sequencing and analysis of a draft version of the human genome was reported by a public consortium (IHGSC, 2001). In the third part

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

of this chapter, we will follow the outline of that 62-page article to describe the human genome from a bioinformatics perspective. We will also describe results described from the Celera Genomics version of the draft human genome sequence (Venter et al., 2001), and subsequent findings.

There are several main starting points for the exploration of the human genome on the web. The NHGRI offers Genome Hub and Ensembl and NCBI (both described below) offer Human Genome Central.

The Genome Hub URL at the National Human Genome Research Institute (NHGRI) at the National Institutes of Health is ▸ http://www.genome.gov/ (then visit "online resources"). This site offers a variety of basic links. Human Genome Central is jointly offered by the Ensembl project (see below; ▸ http://www.ensembl.org/genome/central/) and NCBI (▸ http://www.ncbi.nlm.nih.gov/genome/central). The Oak Ridge National Laboratory (ORNL) offers a human genome website that includes many basic frequently asked questions (▸ http://www.ornl.gov/hgmis/faq/faqs1.html).

Main Conclusions of Human Genome Project

As an introduction to the Human Genome Project, we begin with a summary of its main findings:

1. There are about 30,000–40,000 predicted protein-coding genes in the human genome. This estimate is surprising because we have about the same number of genes as much simpler organisms such as *Arabidopsis thaliana* (26,000 genes) and pufferfish (33,000 genes).
2. The human proteome is far more complex than the set of proteins encoded by invertebrate genomes. Vertebrates have a more complex mixture of protein domain architectures. Additionally, the human genome displays greater complexity in its processing of mRNA transcripts by alternative splicing.
3. Hundreds of human genes were acquired from bacteria by lateral gene transfer, according to the initial report (IHGSC, 2001; Ponting, 2001). Subsequently Salzberg et al. (2001) suggested a revised estimate of 40 genes that underwent horizontal transfer. These genes are homologous to bacterial sequences but appear to lack orthologous genes in other vertebrate and invertebrate species.
4. More than 98% of the human genome does not code for genes. Much of this genomic landscape is occupied by repetitive DNA elements such as long interspersed elements (LINEs) (20%), short interspersed elements (SINEs) (13%), long-terminal-repeat (LTR) retrotransposons (8%), and DNA transposons (3%). Thus half the human genome is derived from transposable elements. However, there has been a decline in the activity of these elements in the hominid lineage. At the same time, the mouse genome displays a continued vigorous activity of transposable elements.
5. Segmental duplication is a frequent occurrence in the human genome, particularly in pericentromeric and subtelomeric regions. This phenomenon is more common than in yeast, fruitfly, or worm genomes. There are three principal ways that gene duplications arise in the human genome (Green and Chakravarti, 2001). First, tandem duplications (created from sequence repeats in a localized region) occur rarely. Second, processed mRNAs are duplicated by retrotransposition. This produces intronless paralogs that are present at one or many sites. Third, segmental duplications occur in which large sections of a chromosome transfer to a new site.
6. There are several hundred thousand *Alu* repeats in the human genome. These have been thought to represent elements that replicate promiscuously. However, their distibution is nonrandom: they are retained in GC-rich regions and thus may confer some benefit on the human genome.

These findings are summarized from several sources, including IHGSC (2001), Venter et al. (2001), and articles by David Baltimore (2001) and Eric Green and Aravinda Chakravarti (2001). Also, the Wellcome Trust Sanger Institute listed "ten facts about the human genome" on its website (▸ http://www.sanger.ac.uk/HGP/publication2001/facts.shtml).

We will define these various types of repetitive elements below.

7. The mutation rate is about twice as high in male meiosis than in female meiosis. This suggests that most mutation occurs in males.
8. More than 1.4 million single-nucleotide polymorphisms (SNPs) were identified. SNPs are single-nucleotide variations that occur once every 100–300 base pairs (bp).

The NCBI database of SNPs currently lists over 4 million RefSNPs (build 115, June 2003; see ▸ http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi). See Chapter 18.

THREE GATEWAYS TO ACCESS THE HUMAN GENOME

The human genome has been sequenced in two parallel projects. A public, international consortium produced a draft sequence in June 2000 that was reported in Feburary 2001 (IHGSC, 2001). That same week, Celera Genomics reported a privately funded effort to sequence the human genome (Venter et al., 2001). As the Celera data are proprietary, most researchers do not have access to the sequence data, the assembly, or the annotation. Thus we will not discuss that project in detail.

The Celera Genomics website is ▸ http://www.celera.com/.

The publicly funded project was assembled and annotated independently several times (Table 17.1). The assembly of the human genome sequence was coordinated by the University of California at Santa Cruz (UCSC, led by David Haussler). The assembly was also performed at NCBI (led by Greg Schuler). The annotation was performed by the Ensembl project (led by Ewan Birney) and separately by Jim Kent and David Haussler (at UCSC) and by the NCBI.

The leader of the project has been Francis Collins, director of the NHGRI. You can read about the NHGRI mission at its website, ▸ http://www.genome.gov/.

Nat Goodman (2002) of the Whitehead Institute/MIT Center for Genome Research discusses his comparison of these resources.

NCBI

The NCBI offers two main ways to access data on the human genome. From the main page of NCBI, you can select "human genome resources," which provides links to each chromosome and a variety of web resources. Alternatively, you can select "human map viewer," which is part of the Entrez genome section (Fig. 17.1). This page allows searches by clicking on a chromosome or by entering a text query.

You can read about the genome resources at NCBI at ▸ http://www.ncbi.nih.gov/About/Doc/hs_genomeintro.html.

The human map viewer integrates human sequence and data from cytogenetic maps, genetic linkage maps, radiation hybrid maps, and YAC chromosomes. A query with "rbp4" (Fig. 17.1) links to the map viewer (Fig. 17.2*a*). This map can display dozens of kinds of information. Some of the options are shown in Figures 17.2*a* (arrow 6) and 17.2*b*) and in Table 17.2. A typical map output (Fig. 17.1) includes a

To access the human map viewer directly, go to ▸ http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/map_search.

TABLE 17-1 Key Websites for Publicly Available Human Genome Sequence Data

Resource	Description	URL
Ensembl	From the EBI and The Wellcome Trust Sanger Institute	▸ http://www.ensembl.org
National Center for Biotechnology Information	Views of chromosomes, maps, and loci	▸ http://ncbi.nlm.nih.gov/genome/guide
University of California at Santa Cruz	Contains an assembly of the draft genome sequence	▸ http://genome.ucsc.edu/

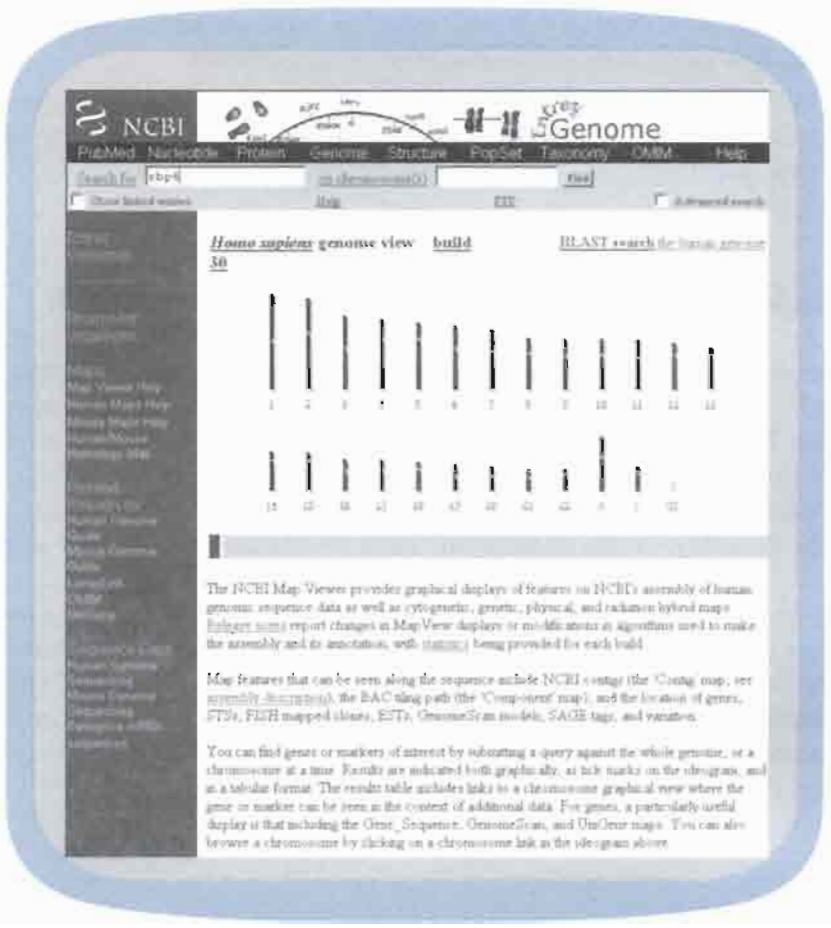


FIGURE 17.1. *The Human Map Viewer is accessible from the main page of NCBI. This resource displays cytogenetic, genetic, physical, and radiation hybrid maps of human genome sequence. Here, entering the query "rbp4" results in links to chromosome 10 (see Fig. 17.2).*

The map viewer provides a link to LocusLink. If instead you begin with any LocusLink entry (Fig. 2.8), you can click "map" to move directly to the map viewer.

variety of links to related resources such as human and mouse UniGene, a human-mouse homology map, and links to genomic contigs that contain the gene in question. Each entry includes "evidence codes" that describe the level of certainty associated with each gene model (Table 17.3). There is also an "evidence viewer" (Fig. 17.3) that displays evidence supporting the existence of a gene and highlights possible discrepancies in the nucleotide sequence, exon–intron boundaries, or other aspects of an annotated gene. As an example, the evidence viewer shows the density of ESTs that have been identified corresponding to each predicted exon of RBP4 (Fig. 17.3).

Ensembl

Ensembl is available at ▸http://www.ensembl.org. The human database is at ▸http://www.ensembl.org/Homo_sapiens/. We described Ensembl projects for mouse, rat, zebrafish, fugu, and mosquito in Chapter 16.

Ensembl is a comprehensive resource for information about the human genome as well as several other genomes (Hubbard et al., 2002). This resource effectively interconnects a wide range of genomics tools with a focus on annotation of known and newly predicted genes. In addition to making annotation information on genes easily accessible, Ensembl provides access to the underlying data that support models of gene prediction. This is described below. The current statistics for the contents of the Ensembl human build are shown in Table 17.4.

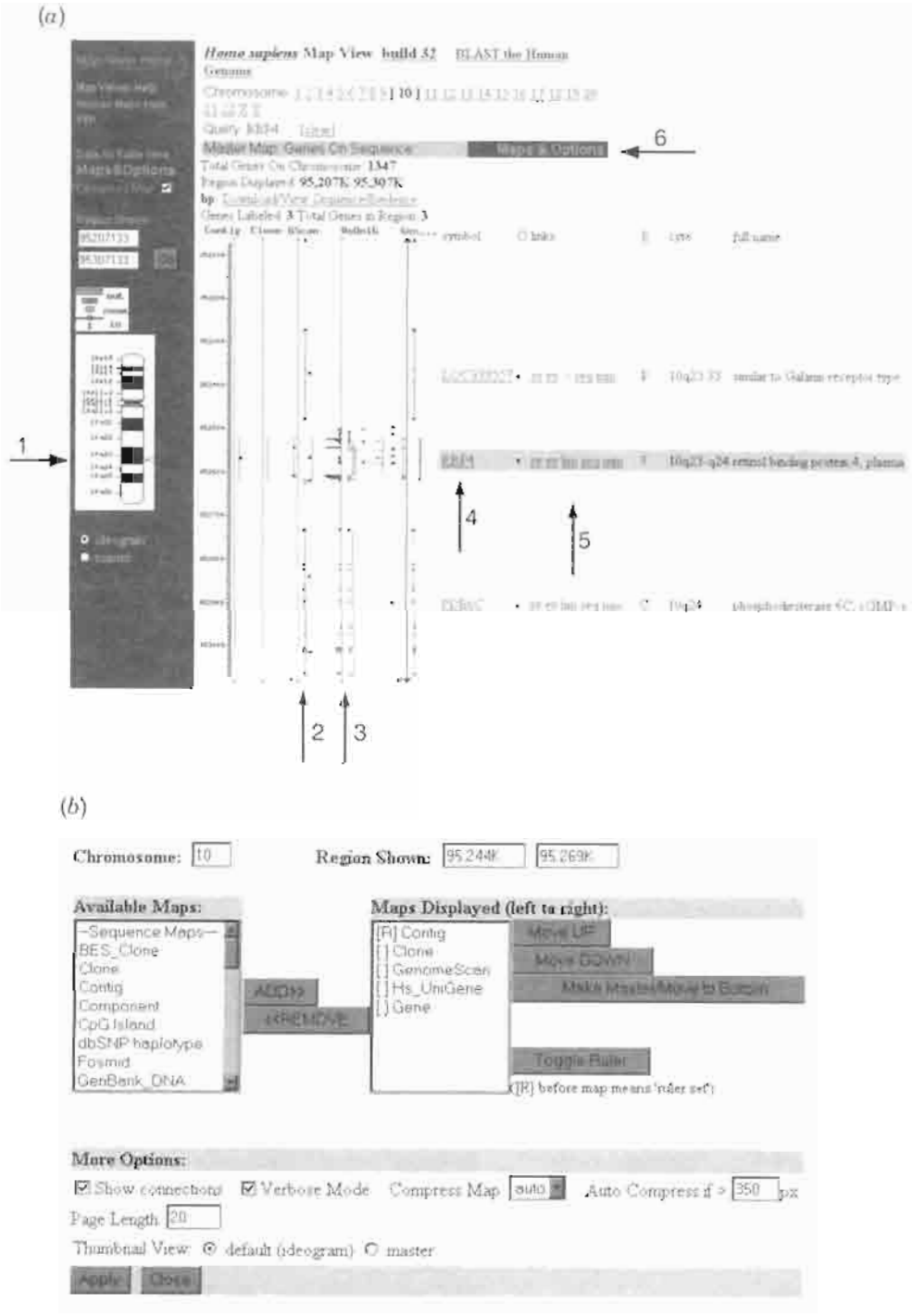


FIGURE 17.2. (a) *The Human Map Viewer from NCBI shows the region of chromosome 10 containing the RBP4 gene. The location is shown on an ideogram (arrow 1), above which is a tool to zoom in or out. The main information that is displayed in column (from left to right across the page) includes a cytogenetic map (arrow 2), a link to the UniGene entry for* RBP4 *(arrow 3), a physical map that can be clicked to a scale of the gene or between 10,000 bp and 10 Mb pairs, and a link to the LocusLink entry (arrow 4). Other links (arrow 5) are sv (Entrez sequence view), ev (Evidence Viewer; Fig. 17.3), hm, seq, and mm. At the top, clicking the display settings (arrow 6) reveals* (b) *a set of pull-down menus. These options allow you to select which fields are displayed in the main graphic in* (a).

TABLE 17-2 User-Selected Map Display Options in NCBI Map Viewer

Cytogenetic maps	Sequence maps
Ideogram	Clone
FISH clone	Contig
Gene_Cytogenetic	Component
Mitelman Breakpoint	CpG island
Morbid/Disease	DbSNP haplotype
Genetic maps	GenBank_DNA
Decode	Gene
Genethon	GenomeScan
Marshfield	SAGE_tag
Radiation hybrid maps	STS
GeneMap99-G3	Transcript (RNA)
GeneMap99-GB4	Hs_UniGene
NCBI RH	Mm_UniGene
Stanford-G3	Variation
TNG	
Whitehead-RH	
Whitehead-YAC	

We saw an example of the Ensembl BLAST server in Figures 5.1 and 5.2.

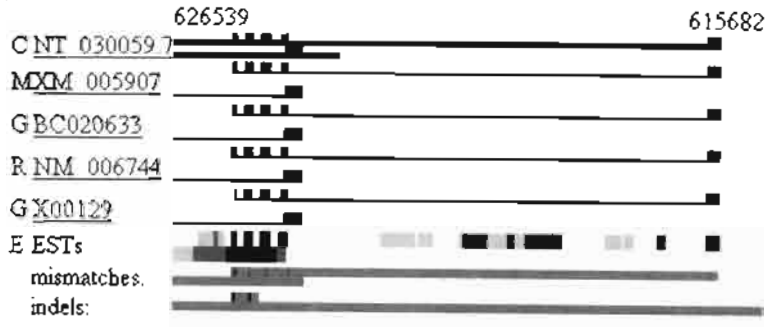
From the main page of Ensembl, you can type a text query (such as RBP4), perform a BLAST search, or browse by chromosome (Fig. 17.4). There are six main entry points to access the Ensembl database:

1. Gene view allows a text query for *RBP4* that leads to a typical Ensembl gene report (Figs. 17.5 and 17.6). This report provides a link to the contig view (Fig. 17.7), "transview" evidence report (Fig. 17.8), protein report (Fig. 17.9), database links (such as RefSeq, Swiss, Prot, LocusLink, and InterPro) and Ensembl homology matches. This view includes the transcript DNA sequence and information on exon–intron boundaries (splice sites). The links to evidence reports are especially important in helping you to evaluate the experimental support for a given gene structure.
2. Contig view allows you to search across an entire chromosome, while also viewing a smaller region in detail. An example is shown in Figure 17.7. The Contig view integrates features from a variety of external data sources such

TABLE 17-3 Evidence Codes for Gene Models on NCBI Map Viewer

Gene Color	Evidence Code	Evidence Used to Construct Gene Model
Blue	C	Confirmed gene model—model based on alignment of mRNA, or mRNAs plus expressed sequence tags (ESTs) to genomic sequence
Light green	E	EST only—model based on EST evidence only
Dark brown	PE	Predicted+EST—model predicted by GenomeScan and EST evidence
Light brown	P	Predicted only—model predicted by GenomeScan
Orange	?	Conflict—there is some discrepancy between the mRNA sequence and the gene model
	I	Interim LocusID—model-based alignment of mRNAs, or mRNAs plus ESTs, to the genome, in which the aligning transcripts could not be unambiguously assigned to a preexisting LocusID

(a)



(b)

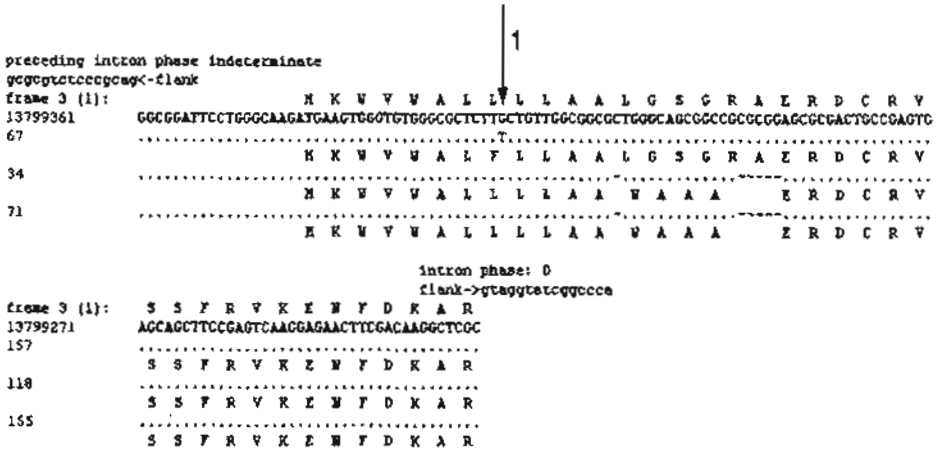


FIGURE 17.3. (a) *The NCBI evidence viewer provides data concerning the structure of human genes. The viewer provides links to the genomic contig (C), the model mRNA from RefSeq (M), a GenBank mRNA (G), a RefSeq mRNA (R), and expressed sequence tags (ESTs) (E) with a color-coding scheme showing the density of ESTs at different positions of the gene. The evidence viewer also describes experimental evidence on each exon.* (b) *For example, in exon 2 of the* RBP4 *gene a nucleotide mismatch was detected (arrow 1) and the exon–intron boundaries are labeled as questionable.*

as UniGene and RefSeq. The tile path shows BAC clones used in the current assembly.

3. Anchor view allows you to select two features from a chromosome as "anchor points" and to display the intervening region.
4. Disease view links to disease entries in OMIM (Chapter 18). In the case of *RBP4*, this view option leads to the relevant OMIM entry for deficiency of RBP4.
5. Map view shows an ideogram of each chromosome, including the known genes, GC content, and SNPs (Fig. 17.10). By clicking on the synteny link (Fig. 17.10, arrow 1), you can see the corresponding region of mouse chromosome 19 where murine *RBP4* is localized (Fig. 17.11).

TABLE 17-4 Statistics for Ensembl Human Web Server (June 2003)

See ▸http://www.ensembl.org/Homo_sapiens/stats/.

Ensembl gene predictions	24,847
Genscan gene predictions	58,770
Ensembl gene exons	248,449
Ensembl gene transcripts	37,347
Contigs	33,840
Clones	26,759
Base pairs	3,242,415,757

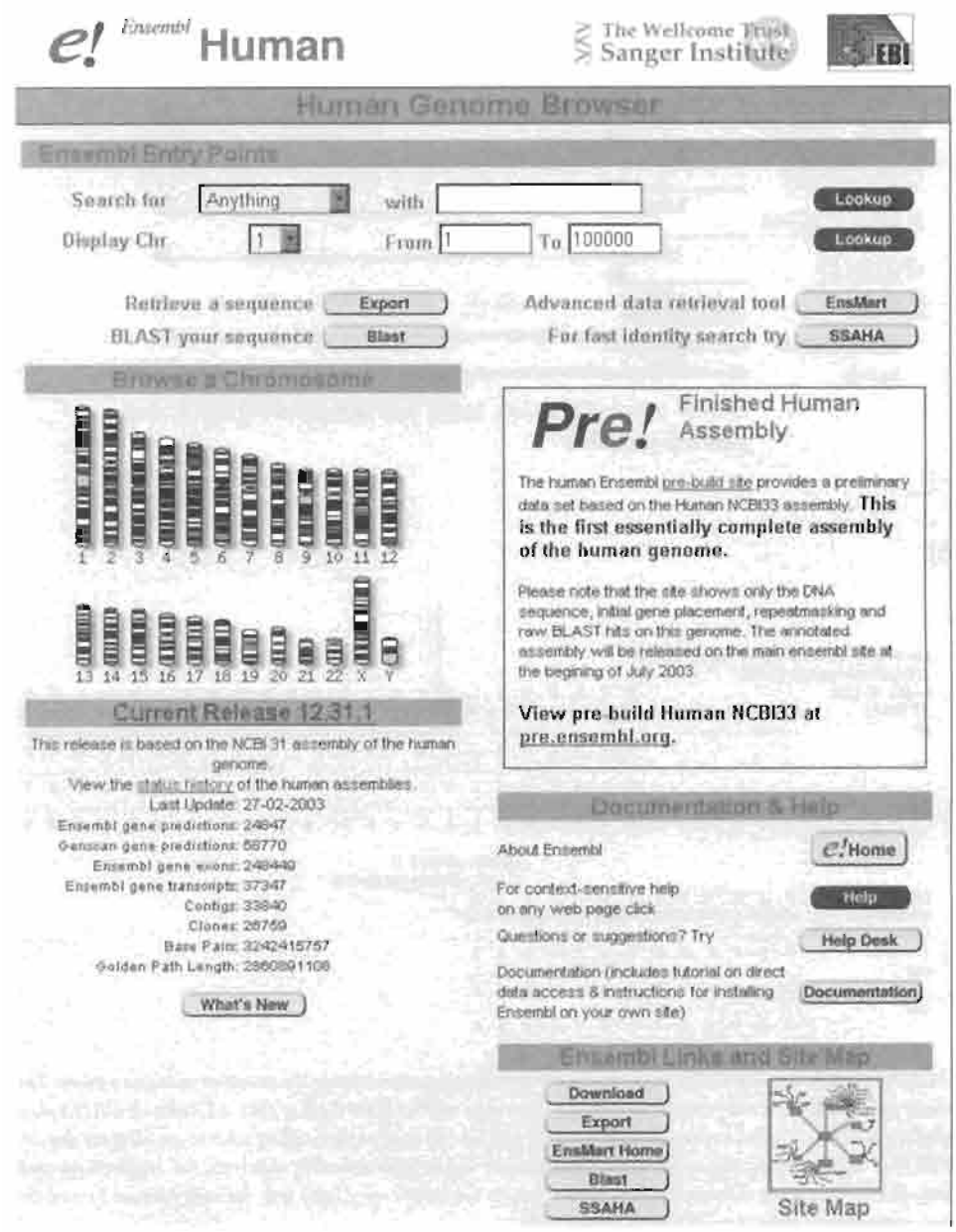


FIGURE 17.4. Front page of Ensembl human genome browser (http://www.ensembl.org/Homo_sapiens/). A direct way to begin searching the site is to enter a search term such as RBP4 (top). The search field can be restricted to a variety of categories using a pull-down menu (e.g., disease, domain, EST, family, gene, mRNA, SNP). Other ways to begin searching include the chromosome ideograms or a BLAST server.

6. Cyto view displays genes, BAC end clones, repetitive elements, and the tiling path across genomic DNA regions.

University of California at Santa Cruz Human Genome Browser

The UCSC Genome Bioinformatics site is accessible at http://genome.ucsc.edu/index.html. It was developed by David Haussler's group (Kent et al., 2002).

The "Golden Path" is the human genome sequence annotated at UCSC. Along with the Ensembl and NCBI sites, the human genome browser at UCSC is one of the three main web-based sources of information for both the human genome and other vertebrate genomes.

The UCSC browser can be searched by keyword, gene name, or other text searches or by a database search with a query using BLAT, the BLAST-like alignment

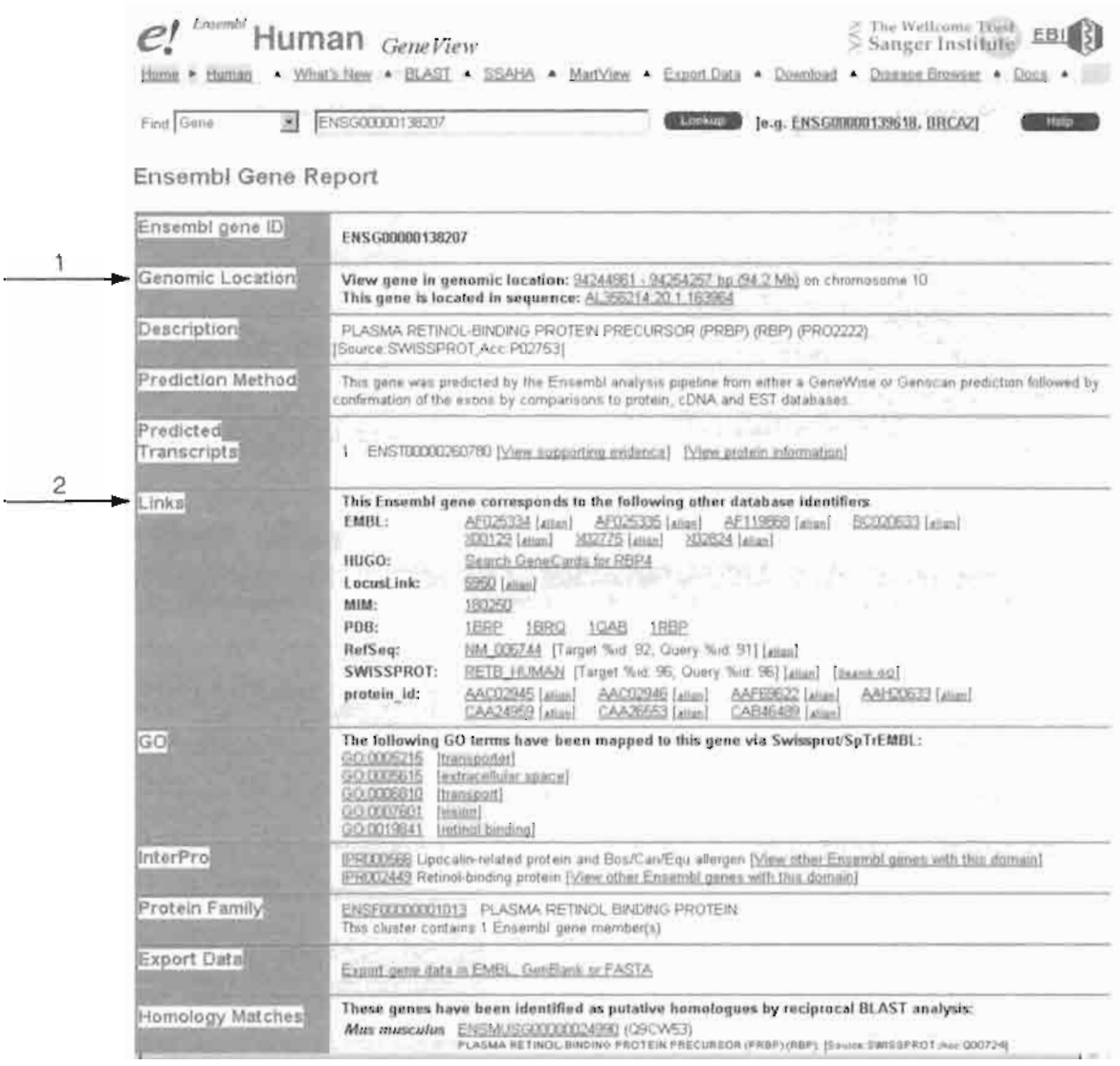


FIGURE 17.5. *A typical Ensembl gene report (top portion only) is shown for human RBP4. This contains information including a link to the genomic contig (arrow 1; Fig. 17.7, below) and links to a variety of databases we have studied already (arrow 2).*

tool (Kent, 2002) (see Chapter 5). A search for retinol-binding protein ("rbp4") produces the output shown in Figure 17.12. This includes a top section where the DNA scale can be changed, and the user can scroll across a chromosome. The main output window includes rows of annotation on the *RBP4* gene. At the bottom, dozens of additional annotation tracks can be selected. These include the following (Kent et al., 2002):

The UCSC site is based on a human genome assembly provided by NCBI

- Assembly contigs and gaps
- Messenger RNA and EST alignments and gene expression data from SAGE or microarrays
- Multiple gene predictions; in the case of *RBP4,* not all predictions of exons are consistent
- Cross-species homologies

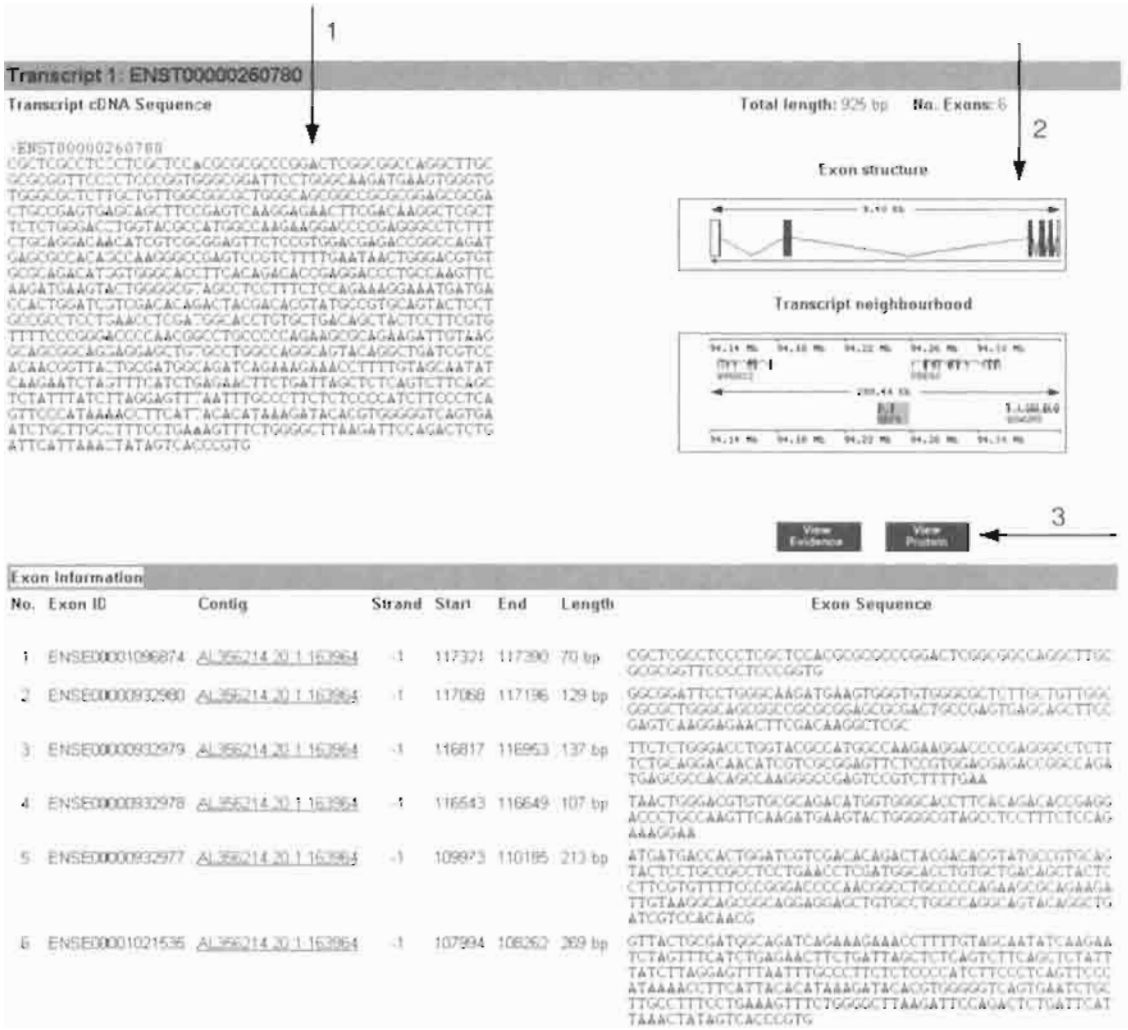


FIGURE 17.6. *The middle portion of a typical Ensembl gene report is shown for human RBP4. The RBP4 cDNA sequence is given (arrow 1) as well as a graphical map of the exon structure of the gene (arrow 2) and neighboring genes. The "view evidence" and "view protein" links (arrow 3) provide data shown in Figures 17.8 and 17.9. The bottom of the Ensembl gene report includes the exons and the predicted splice sites (not shown).*

- SNPs
- Repetitive DNA
- Radiation hybrid data

An example of the use of an expanded annotation field is shown in Figure 17.13. Data are included for nonhuman mRNAs that match *RBP4,* mouse, chimpanzee, and *Tetraodon nigriviridis* alignments from BLAT, and gene expression data for *RBP4* from a collection of 60 cell lines for which 8000 expressions were measured for 8000 genes (Ross et al., 2000).

In addition to these annotations, this genome browser allows users to add customized annotations of any data. Approximately 50% of the annotation tracks were calculated at UCSC, while the other half were provided by researchers who use this site (Kent et al., 2002).

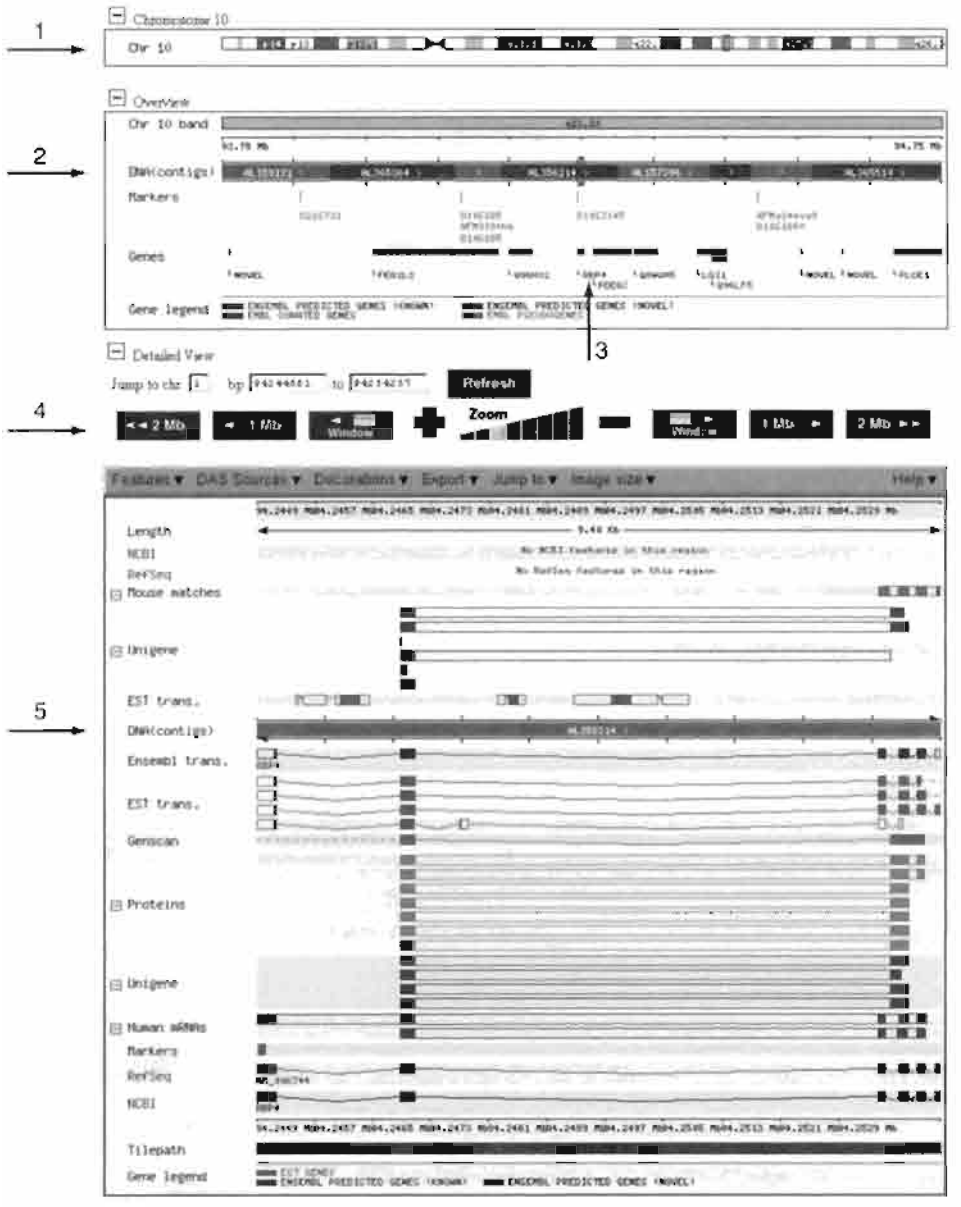


FIGURE 17.7. *The Ensembl ContigView (linked from the* RBP4 *gene report, Fig. 17.5) shows an ideogram of chromosome 10 (arrow 1), including the position of the* RBP4 *gene (boxed). The overview section includes an overview of contigs in the 10q23.33 region (arrow 2) and indicates annotated genes including* RBP4 *(arrow 3). The detailed view section can be zoomed and scrolled (arrow 4). It includes a schematic view of a contig (arrow 5) with the annotations on the top strand above it and the annotations on the bottom strand below (including* RBP4 *links to ESTs, markers, and predicted transcripts).*

HUMAN GENOME PROJECT

The two articles on the human genome project that appeared in February 2001 provide an initial glimpse of the genome (IHGSC, 2001; Venter et al., 2001). In the remainder of this chapter, we will follow the outline of the public consortium paper (IHGSC, 2001). We will not summarize all the major findings, but we will focus on selected topics. While the reported sequence represents 90% completion of the human genome, additional publications have described the sequence of human chromosomes 20, 21, and 22 in detail (Dunham et al., 1999; Deloukas et al., 2001; Hattori et al., 2000). The finished sequence is complete with an accuracy of at least 99.99% and no gaps. These chromosomal sequences have few gaps and more than 95% coverage of the euchromatin, thus meeting the goals of the Human Genome Project.

The euchromatin is the gene-containing part of the genome.

FIGURE 17.8. *The Ensembl gene report (Fig. 17.6, arrow 3) includes a link to evidence for the existence of a particular transcript, shown here. This resource includes graphical evidence for the presence of each of the six exons in the* RBP4 *gene. The presence of each exon is tabulated from many data libraries (only several are shown here).*

Background of Human Genome Project

The National Academy Press (▸ http://www.nap.edu) offers this 1988 book free on-line at ▸ http://www.nap.edu/books/0309038405/html/.

The Human Genome Project was first proposed by the U.S. National Research Council (1988). This report proposed the creation of genetic, physical, and sequence maps of the human genome. At the same time, parallel efforts were supported for model organisms (bacteria, *Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster,* and *Mus musculus*).

The major goals of the Human Genome Project are listed in Table 17.5.

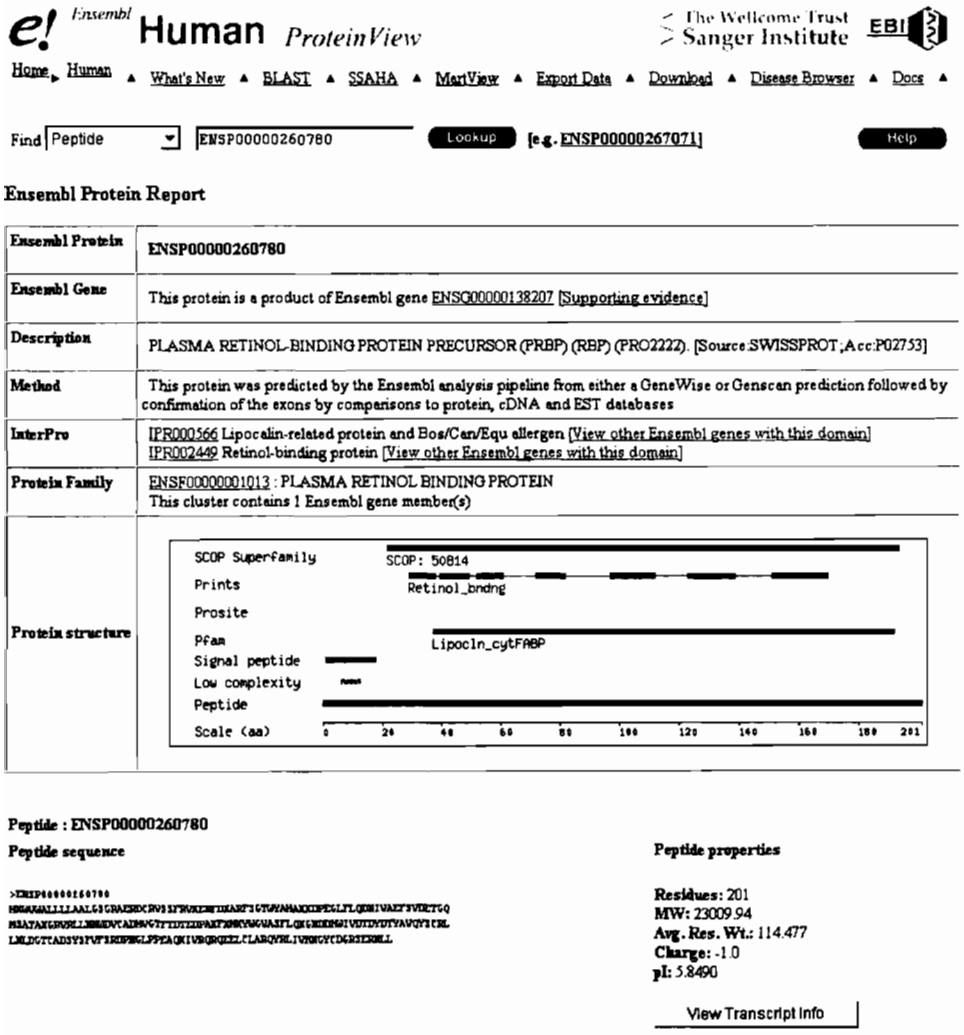


FIGURE 17.9. *The Ensembl protein report for RBP4 (see Fig. 17.6, arrow 3) shows features such as the predicted protein structure, the amino acid sequence, and the domains present in PRINTS, Pfam, and other protein databases.*

One component of the human genome project is the Ethical, Legal and Social Issues (ELSI) initiative. From 3 to 5% of the annual budget has been devoted to ELSI, making it the world's largest bioethics project.

You can read about ELSI at ▸http://www.ornl.gov/hgmis/elsi/elsi.html.

Examples of issues addressed by ELSI include:

- Who owns genetic information?
- Who should have access to genetic information?
- How does genomic information affect members of minority communities?
- What societal issues are raised by new reproductive technologies?
- How should genetic tests be regulated for reliability and validity?
- To what extent do genes determine behavior?
- Are there health risks associated with genetically modified foods?

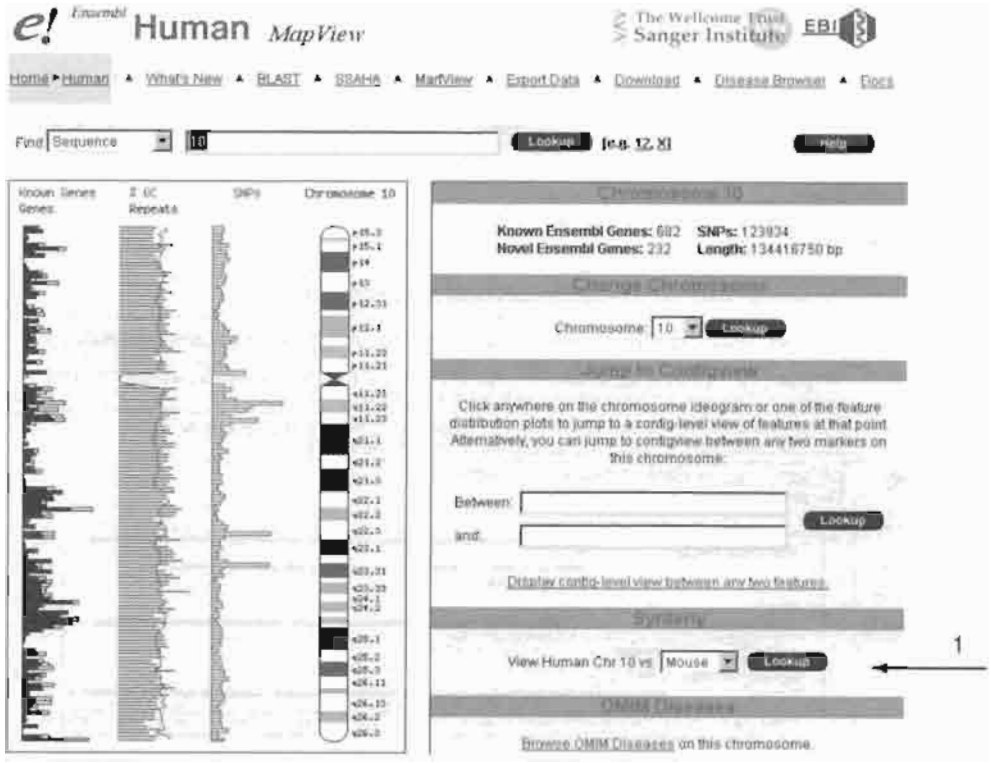


FIGURE 17.10. *The Ensembl MapView displays each chromosome with associated data on genes, SNPs, and GC content. There is a link to conserved syntenic (corresponding) regions in the mouse genome (arrow 1) (see Fig. 17.11).*

Strategic Issues: Hierarchical Shotgun Sequencing to Generate Draft Sequence

The public consortium approach to sequencing the human genome was to employ the hierarchical shotgun sequencing strategy. The rationale for taking this approach was as follows:

- Shotgun sequencing can be applied to DNA molecules of many sizes, including plasmids (typically several kilobases), cosmid clones [40 kilobases (kb)], yeast, and BACs [up to 1 or 2 megabases (mb)].
- The human genome has large amounts of repetitive DNA (about 50% of the genome; see below). Whole-genome shotgun sequencing, the main approach taken by Celera Genomics, was not adopted by the public consortium because of the difficulties associated with assembling repetitive DNA fragments. In the public consortium approach, large-insert clones (typically 100–200 kb) from defined chromosomes were sequenced.
- The reduction of the sequencing project to specific chromosomes allowed the international team to reduce and distribute the sequencing project to a set of sequencing centers. These centers are listed in Table 17.6.

Early in the evolution of the Human Genome Project, it was thought that breakthroughs in DNA sequencing technology would be necessary to allow the completion of such a large-scale project. This did not occur. Instead, the basic principles

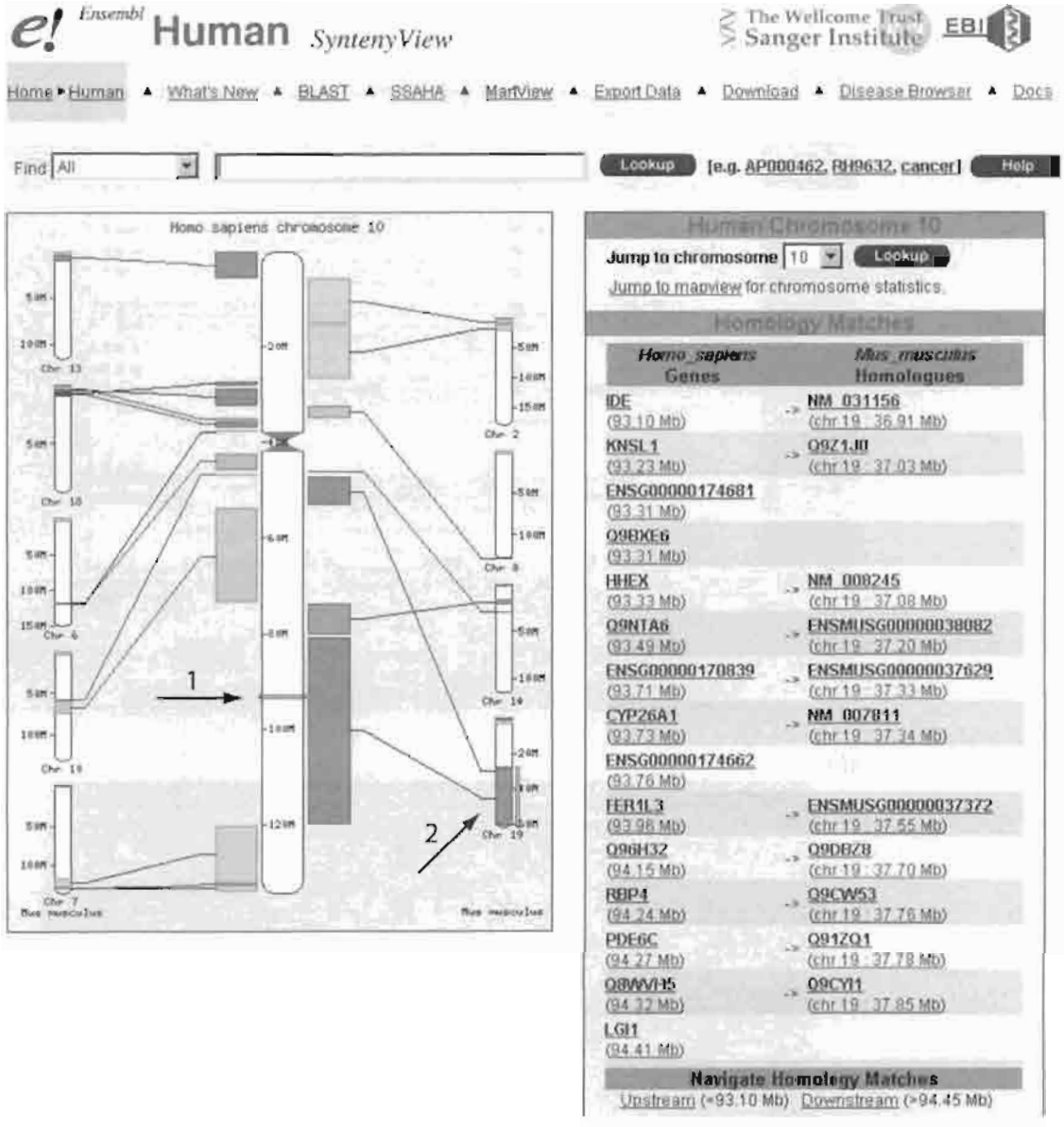


FIGURE 17.11. *The MapView at Ensembl links to human/mouse conserved syntenic maps. This map shows human chromosome 10, including the region with the* RBP4 *gene (arrow 1 and gene list at right) as well as the corresponding region of mouse chromosome 19 (arrow 2).*

of dideoxynucleotide sequencing by the method of Sanger (see Chapter 12) were improved upon. Some recent innovations include (reviewed by Green, 2001):

- Advances in the automatic detection of DNA molecules, such as capillary electrophoresis–based sequencing machines
- Improved thermostable polymerases
- Fluorescent dye-labeled dideoxynucleotide terminators

The most important of these advances was probably the introduction of the automated sequencer in the late 1980s. The public consortium draft genome sequence was generated by selecting BAC clones to be sequenced, sequencing the clones, and assembling the clones. Most libraries contained BAC clones or P1-derived artificial

The October 2000 draft version of the human genome was based on the sequence and assembly of over 29,000 BAC clones with a total length of 4.26 billion base pairs (Gb). There were 23 Gb of raw shotgun sequence data.

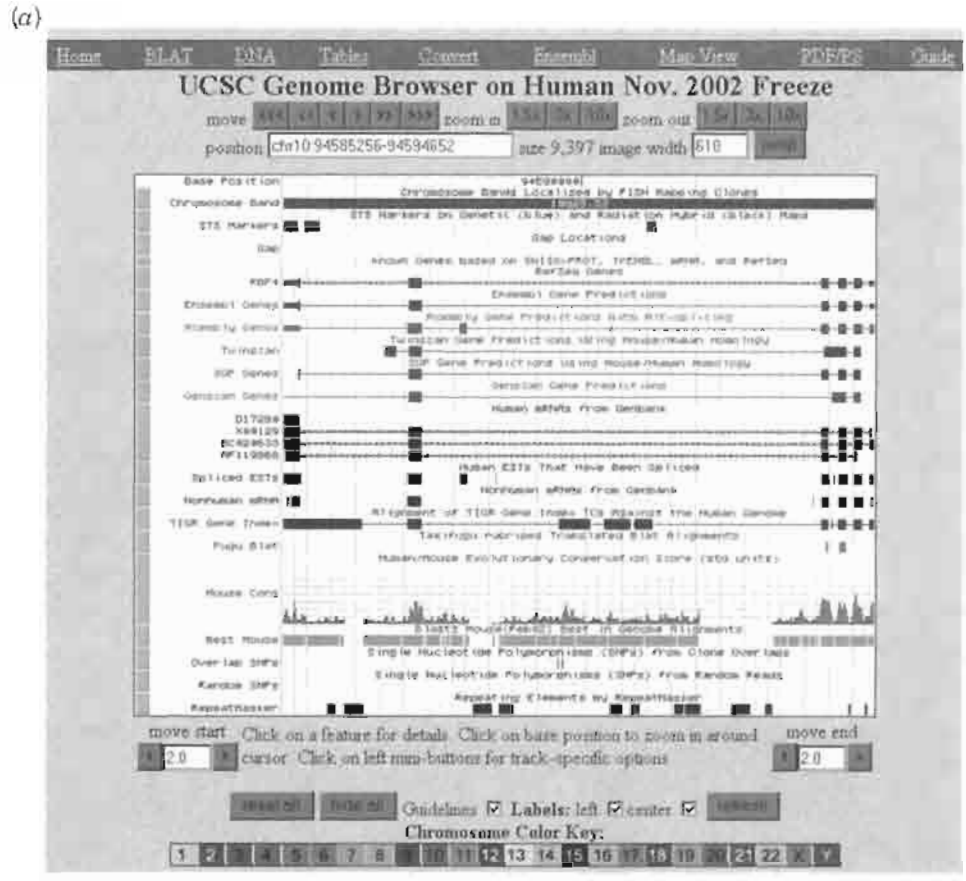


FIGURE 17.12. *The human genome browser at UCSC (http://genome.ucsc.edu) (Kent et al., 2002). At top, the user can scroll across a chromosome. In the middle, genome annotations are displayed. At the bottom, a variety of track controls allow the display to be dynamically modified with dozens of types of annotation such as repetitive DNA, comparative genomics, and gene expression data.*

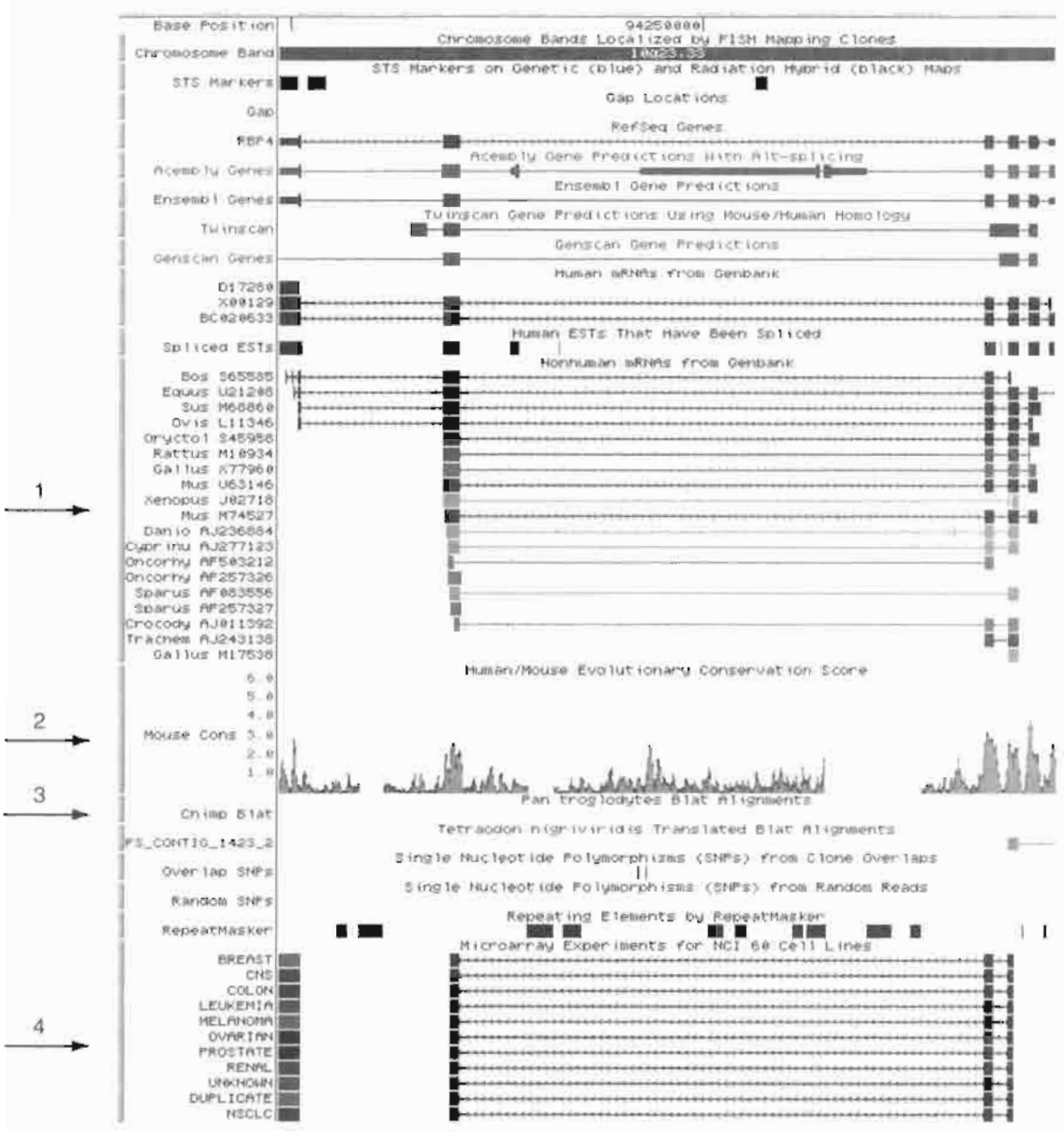


FIGURE 17.13. *The UCSC human genome browser provides annotations that are easily modifiable. Here, the* RBP4 *entry (Fig. 17.12) is expanded to show exons with matches in nonhuman species (arrow 1), mouse/human conserved regions (arrow 2), chimpanzee matches (arrow 3), and expression data from the "NCI60" experiment of the expression of 8000 genes in 60 cell lines (this shows that* RBP4 *is up regulated in a cancer cell line) (arrow 4).*

clones (PACs). These libraries were prepared from DNA obtained from anonymous donors. Selected clones were subjected to shotgun sequencing.

In conjunction with sequencing of BAC and other large-insert clones, the sequence data were assembled into an integrated draft sequence (Table 17.7). An example of the procedure is shown in Figure 17.14.

The NCBI Contig Assembly and Annotation Process is described at ▸http://www.ncbi.nlm.nih.gov/genome/guide/build.html.

Features of Genome Sequence

A draft genome sequence contains a mixture of finished, draft, and predraft data. A key aspect of the sequence is the extent to which the sequenced fragments are

TABLE 17-5 Eight Goals of Human Genome Project (1998–2003)

1. Human DNA sequence	• Finish the complete human genome sequence by the end of 2003. • Achieve coverage of at least 90% of the genome in a working draft based on mapped clones by the end of 2001. • Make the sequence totally and freely accessible.
2. Sequencing technology	• Continue to increase the throughput and reduce the cost of current sequencing technology. • Support research on novel technologies that can lead to significant improvements in sequencing technology. • Develop effective methods for the development and introduction of new sequencing technologies.
3. Human genome sequence variation	• Develop technologies for rapid, large-scale identification and/or scoring of single-nucleotide polymorphisms and other DNA sequence variants. • Identify common variants in the coding regions of the majority of identified genes during this five-year period. • Create a SNP map of at least 100,000 markers. • Create public resources of DNA samples and cell lines.
4. Functional genomics technology	• Generate sets of full-length cDNA clones and sequences that represent human genes and model organisms. • Support research on methods for studying functions of nonprotein-coding sequences. • Develop technology for comprehensive analysis of gene expression. • Improve methods for genomewide mutagenesis. • Develop technology for large-scale protein analyses.
5. Comparative genomics	• Complete the sequence of the roundworm *C. elegans* genome and the fruitfly *Drosophila* genome • Develop an integrated physical and genetic map for the mouse, generate additional mouse cDNA resources, and complete the sequence of the mouse genome by 2008.
6. Ethical, legal, and social issues	• Examine issues surrounding completion of the human DNA sequence and the study of genetic variation. • Examine issues raised by the integration of genetic technologies and information into health care and public health activities. • Examine issues raised by the integration of knowledge about genomics and gene–environment interactions in nonclinical settings. • Explore how new genetic knowledge may interact with a variety of philosophical, theological, and ethical perspectives. • Explore how racial, ethnic, and socioeconomic factors affect the use, understanding, and interpretation of genetic information, the use of genetic services, and the development of policy.
7. Bioinformatics and computational biology	• Improve content and utility of databases. • Develop better tools for data generation, capture, and annotation. • Develop and improve tools and databases for comprehensive functional studies. • Develop and improve tools for representing and analyzing sequence similarity and variation. • Create mechanisms to support effective approaches for producing robust, exportable software that can be widely shared.
8. Training and manpower	• Nurture the training of scientists skilled in genomics research. • Encourage the establishment of academic career paths for genomic scientists. • Increase the number of scholars who are knowledgeable in both genomic and genetic sciences and in ethics, law, or the social sciences.

Source: Adapted from ▸ http://www.ornl.gov/TechResources/Human_Genome/hg5yp/goal.html.

TABLE 17-6 Twenty Institutions That Form Human Genome Sequencing Consortium

Genome Sequencing Center	Location/Description	URL
Baylor College of Medicine	Houston, Texas	http://www.hgsc.bcm.tmc.edu/
Beijing Human Genome Center, Institute of Genetics, Chinese Academy of Sciences	Beijing, China	http://hgc.igtp.ac.cn/
Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center	Cold Spring Harbor, New York	http://nucleus.cshl.org/ genseq/lita_annenberg_hazen_genome_cent.htm
Gesellschaft fur Biotechnologische Forschung mbH	Braunschweig, Germany	http://genome.gbf.de/
Genoscope	Evry, France	http://www.genoscope.cns.fr/externe/English/Projets/projets.html
Genome Therapeutics Corporation	Waltham, MA	http://www.genomecorp.com/
Institute for Molecular Biotechnology	Jena, Germany	http://genome.imb-jena.de/
Joint Genome Institute, U.S. Department of Energy	Walnut Creek, California	http://www.jgi.doe.gov/
Keio University	Tokyo, Japan	http://www-alis.tokyo.jst.go.jp/HGS/top.pl
Max Planck Institute for Molecular Genetics	Berlin, Germany	http://seq.mpimg-berlin-dahlem.mpg.de/
Multimegabase Sequencing Center, Institute for Systems Biology	Seattle, Washington	http://www.systemsbiology.org/
RIKEN Genomic Sciences Center	Saitama, Japan	http://hgp.gsc.riken.go.jp/
The Sanger Centre	Hinxton, United Kingdom	http://www.sanger.ac.uk/HGP/
Stanford Genome Technology Center	Palo Alto, California	http://www-sequence.stanford.edu/
Stanford Human Genome Center	Palo Alto, California	http://shgc.stanford.edu/
University of Oklahoma's Advanced Center for Genome Technology	Norman, Oklahoma	http://www.genome.ou.edu/
University of Texas Southwestern Medical Center	Dallas, Texas	http://www3.utsouthwestern.edu/ index.htm
University of Washington Genome Center	Seattle, Washington	http://www.genome.washington.edu/UWGC/
Washington University Genome Sequencing Center	St. Louis, Missouri	http://genome.wustl.edu/gsc/
Whitehead Institute for Biomedical Research, MIT	Cambridge, Massachusetts	http://www-genome.wi.mit.edu/

contiguous. The average length of a clone or a contig is not a consistently useful measure of the extent to which a genome has been sequenced an assembled. Instead the N50 length describes the largest length L such that 50% of all nucleotides are contained in contigs or scaffolds of at least size L. For the draft version of the human genome sequence, half of all nucleotides are present in a fingerprint clone contig of at least 8.4 Mb (Table 17.8).

N50 statistics are reported at the UCSC genome browser (http://genome.ucsc.edu/goldenPath/stats.html).

The quality of the genome sequence is assessed by counting the number of gaps and by measuring the nucleotide accuracy. About 91% of the unfinished draft sequence had an error rate of less than 1 per 10,000 bases (PHRAP score >40).

A PHRAP score of 40 corresponds to an error probability of $10^{-40/10}$, or 99.99% accuracy (see Chapter 12).

TABLE 17-7 Contigs Categorized by Size from NCBI Build 30 (November 2002)

See ▸ http://www.ncbi.nlm.nih.gov/genome/guide/human/HsStats.html.

Range (kb)	Number	Length (kb)	Percent of Total
<30	30	469	0.01
30–100	113	7,088	0.24
100–250	547	92,587	3.21
250–500	218	78,223	2.71
500–1000	227	167,372	5.81
1000–5000	479	1,079,538	37.52
>5000	122	1,451,804	50.46

Another fundamental category of error is misassembly. This can be especially problematic for regions of highly repetitive DNA or for duplicated regions of the genome (see below; Eichler, 2001). Comparisons of the genome sequences produced by the IHGSC and Celera Genomics indicate substantial differences that may reflect differences in assembly (Li et al., 2002).

The sizes of the human chromosomes are listed in Table 17.9. The autosomes are numbered approximately in order of size; the first to be completely sequenced were chromosomes 22 (Dunham et al., 1999), 21 (Hattori et al., 2000), and 20 (Deloukas et al., 2001). These are among the smallest chromosomes.

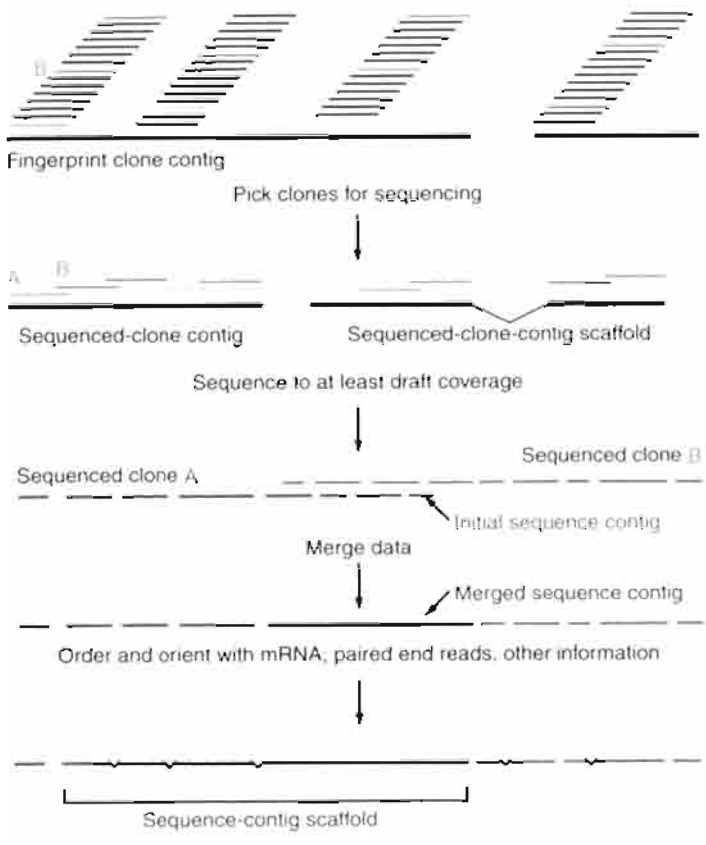


FIGURE 17.14. *Clone and sequence coverage of the human genome. A fingerprint clone contig is assembled based on restriction endonuclease digestion patterns in order to select clones that are inferred to overlap for sequencing. These sequenced-clone contigs are merged to generate scaffolds in which the order and orientation of each clone is established. (From IHGSC, 2001.) Used with permission.*

Broad Genomic Landscape

Having a nearly complete view of the nucleotide sequence of the human genome, we can explore its broad features. These include:

- The distribution of GC content
- CpG islands and recombination rates
- The repeat content
- The gene content

We will next examine each of these four features of the genome. Using the genome browsers at NCBI, Ensembl, and UCSC, we can explore the genomic landscape from the level of single nucleotides to entire chromosomes.

Long-Range Variation in GC Content

The average GC content of the human genome is 41%. However, there are regions that are relatively GC rich and GC poor. A histogram of the overall GC content (in

TABLE 17-8 Continuity of Draft Genome Sequence

This is described by N50 statistics, which report the length of various clone types for which 50% of the nucleotides reside.

Clone	Length (L)
Initial sequencing contig	21.7 kb
Sequence contig	82 kb
Sequence-contig scaffold	274 kb
Sequenced-clone contig	826 kb
Fingerprint clone contig	8.4 Mb

Source: Adapted from IHGSC (2001).

TABLE 17-9 Sizes of Human Chromosomes

Size refers to euchromatic size of the chromosome.

Chromosome	Size (Mb)	Number of Contigs	Finished (%)
1	263	397	32
2	255	436	55
3	214	147	20
4	203	247	21
5	194	431	40
6	183	131	86
7	171	265	79
8	155	85	13
9	145	206	47
10	144	248	47
11	144	183	43
12	143	283	53
13	98	107	84
14	93	69	86
15	89	64	26
16	98	156	34
17	92	138	38
18	85	59	15
19	67	185	61
20	72	9	82
21	34	5	99.5
22	34	12	97
X	164	257	71
Y	35	6	64
all	3175		

Source: Adapted from IHGSC (2001) and NCBI (http://www.ncbi.nlm.nih.gov/genome/guide/H_sapiens.html, October 2001).

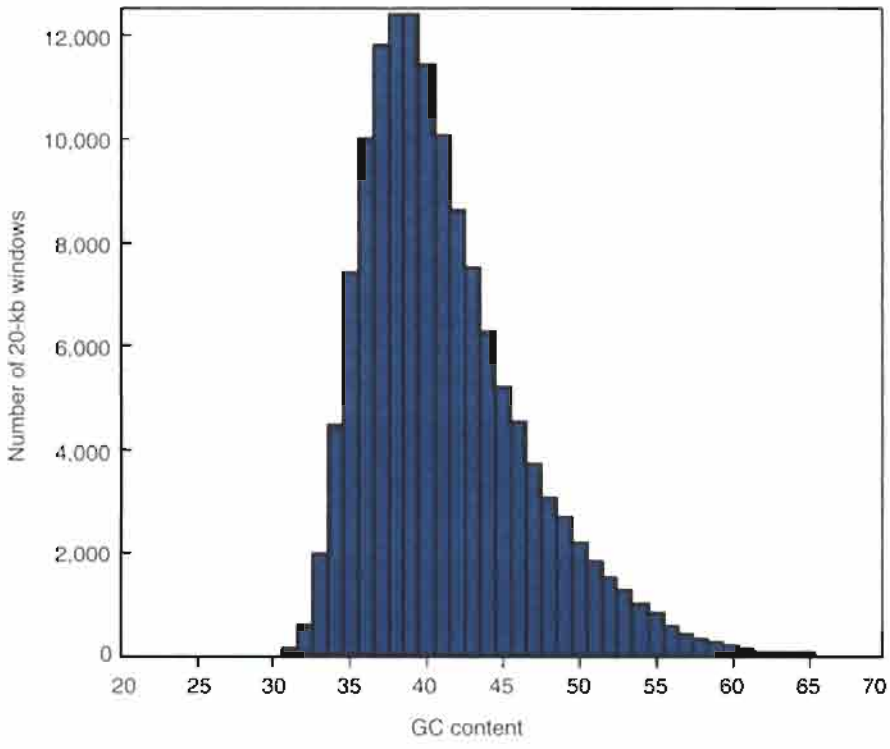


FIGURE 17.15. *Histogram of percent GC content versus the number of 20-kb windows in the draft human genome sequence. Note that the distribution is skewed to the right, with a mean GC content of 41%. (From IHGSC, 2001.) Used with permission.*

You can view GC content across any chromosome in the NCBI, Ensembl, or UCSC genome browsers. For example, in the Ensembl browser (Fig. 17.7) click "decorations" to add a GC content layer.

The L (light) and H (heavy) designations for isochores refer to the sedimentation behavior of genomic DNA in cesium chloride gradients. Genomic DNA fragments migrate to different positions based on their percent GC content.

Gene silencing refers to transcriptional repression. We briefly described MeCP2, a protein that binds to methylated CpG islands (Fig. 8.3). MeCP2 further recruits proteins such as a histone deactylase that alters chromatin structure and represses transcription. Mutations in *MECP2*, the X-linked gene encoding MeCP2, cause Rett sydrome (Amir et al., 1999). This disease causes distinctive neurological symptoms in girls, including loss of purposeful hand movements, seizures, and autistic-like behavior (Chapter 18).

20-kb windows) shows a broad profile with skewing to the right (Fig. 17.15). Fifty-eight percent of the GC content bins are below the average, while 42% are above the average, including a long tail of highly GC-rich regions.

Giorgio Bernardi and colleagues have proposed that mammalian genomes are organized into a mosaic of large DNA segments (e.g., >300 kb) called isochores. These isochores are fairly homogeneous compositionally and can be divided into GC-poor families (L1 and L2) or GC-rich families (H1, H2, and H3). The IHGSC (2001) report did not identify clearly defined isochores, and Haring and Kypr (2001) did not detect isochores in human chromosomes 21 and 22. Subsequent analyses by Bernardi and colleagues (Bernardi, 2001; Pavlicek et al., 2002) do support the mosaic organization of the human genome by GC content. The discrepancies depend in part on the size of the window of genomic DNA that is analyzed.

CpG Islands

The dinucleotide CpG is greatly underrepresented in genomic DNA, occurring at about one-fifth its expected frequency. Most CpG dinucleotides are methylated on the cytosine and subsequently are deaminated to thymine bases. However, the genome contains many "CpG islands" which are typically associated with the promoter and exonic regions of housekeeping genes (Gardiner-Garden and Frommer, 1987). CpG islands have roles in processes such as gene silencing, genomic imprinting (Tycko and Morison, 2002), and X-chromosome inactivation (Avner and Heard, 2001).

You can display predicted CpG islands in genomic DNA at the NCBI, Ensembl, and UCSC genome browser websites (e.g., Table 17.2 and Fig. 17.12). According to the IHGSC (2001), there are 50,267 predicted CpG islands in the human genome. After blocking repetitive DNA sequences with RepeatMasker, there were 28,890 CpG islands. (This lower number reflects the high GC content of *Alu* repeats as seen

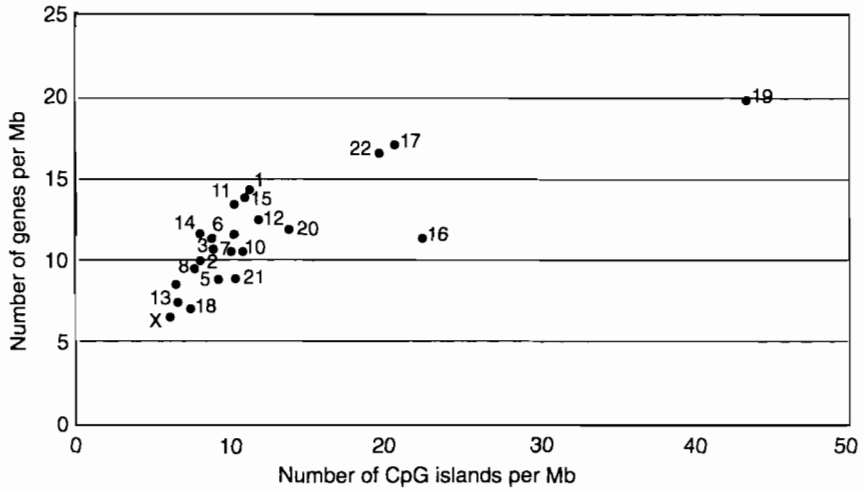


FIGURE 17.16. *The number of CpG islands per megabase is plotted versus the number of genes per megabase as a function of chromosome. Note that chromosome 19, the most gene-rich chromosome, has the greatest number of CpG islands per megabase. (From IHGSC, 2001.) Used with permission.*

in Fig. 16.8.) There are 5–15 CpG islands per megabase of DNA on most chromosomes, although chromosome 19 (the most gene-dense chromosome) contains 43 CpG islands per megabase (Fig. 17.16).

Genomic imprinting is the differential expression of genes from maternal and paternal alleles. Tycko and Morison (2002) offer a database of imprinted genes (▸http://www.otago.ac.nz/IGC).

X-chromosome inactivation is a dosage compensation mechanism in which cells in a female body selectively silence the expression of genes from either the maternally or paternally derived X chromosome (Avner and Heard, 2001).

Comparison of Genetic and Physical Distance

It is possible to compare the genetic maps and physical maps of the chromosomes to estimate the rate of recombination per nucleotide (Yu et al., 2001). Genetic maps, also known as linkage maps, are chromosome maps based on meiotic recombination. During meiosis the two copies of each chromosome present in each cell are reduced to one. The homologous parental chromosomes recombine (exchange DNA) during this process. Genetic maps describe the distances between DNA sequences (genes) based on their frequency of recombination. Thus genetic maps describe DNA sequences in units of centimorgans (cM), which describe relative distance. One centimorgan corresponds to 1% recombination.

In contrast to genetic maps, physical maps describe the physical position of nucleotide sequences along each chromosome. With the completion of draft versions of the human genome, it became possible to compare genetic and physical maps.

Figure 17.17 shows a plot of genetic distance (y axis; in centimorgans) versus physical distance for human chromosome 12 (x axis; in megabases) (IHGSC, 2001). There are two main conclusions. First, the recombination rate tends to be suppressed near the centromeres (note the flat slope in Fig. 17.17, arrow 1), while the recombination rate is far higher near the telomeres. This effect is especially pronounced in males. Second, long chromosome arms tend to have an average recombination rate of 1 cM/Mb, while the shortest arms have a much higher average recombination rate (above 2 cM/Mb). The range of the recombination rate throughout the genome varies from 0 to 9 cM/Mb (Yu et al., 2001). These researchers identified 19 recombination "deserts" (up to 5 Mb in length with sex-average recombination rates below 0.3 cM/Mb) and 12 recombination "jungles" (up to 6 Mb in length with sex-average recombination rates above 3.0 cM/Mb).

The NCBI, Ensembl, and UCSC genome browsers allow you to view both genetic and physical maps. For example, NCBI offers three kinds of genetic maps in its map viewer (Table 17.2 and Fig. 17.2).

Repeat Content of Human Genome

Repetitive DNA probably occupies over 50% of the human genome. The origin of these repeats and their function present fascinating questions. What different kinds

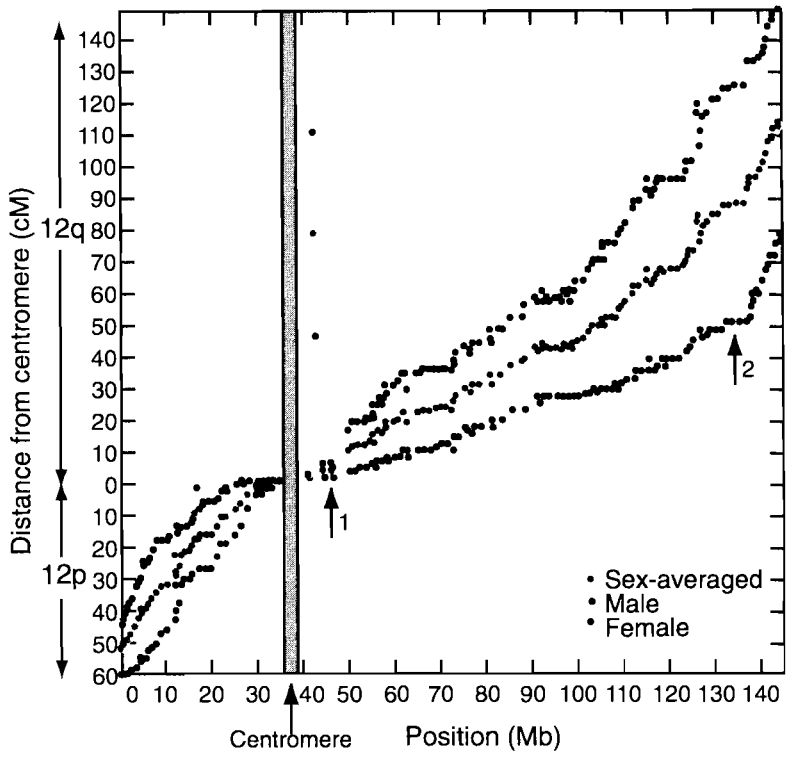


FIGURE 17.17. *Comparison of physical distance (in megabases, x axis) with genetic distance (in centimorgans, y axis) for human chromosome 12. Note that the recombination rate tends to be lower near the centromere (arrow 1) and higher near the telomeres (distal portion of each chromosome). The recombination is especially high in the male meiotic map (arrow 2). (From IHGSC, 2001.) Used with permission.*

of repeats occur? From where did they originate and when? Is there a logic to their promiscuous growth in our genomes or do they multiply without purpose? One of the outcomes of the Human Genome Project is that we are beginning to understand the extent and nature of the repeat content of our genome.

There are five main classes of repetitive DNA in humans (IHGSC, 2001; Jurka, 1998):

1. Interspersed repeats (transposon-derived repeats)
2. Processed pseudogenes: inactive, partially retroposed copies of protein-coding genes
3. Simple sequence repeats: microsatellites and minisatellites, including short sequences such as $(A)_n$, $(CA)_n$, or $(CGG)_n$
4. Segmental duplications, consisting of blocks of 10–300 kb that are copied from one genomic region to another
5. Blocks of tandemly repeated sequences such as are found at centromeres, telomeres, and ribosomal gene clusters.

We will briefly explore each of these repeats.

Transposon-Derived Repeats

Incredibly, 45% of the human genome or more consists of repeats derived from transposons. These are often called interspersed repeats. Many transposon-derived repeats replicated in the human genome in the distant past (hundreds of millions of years ago), and thus because of sequence divergence it is possible that the 45% value is an underestimate. Transposon-derived repeats can be classified in four categories

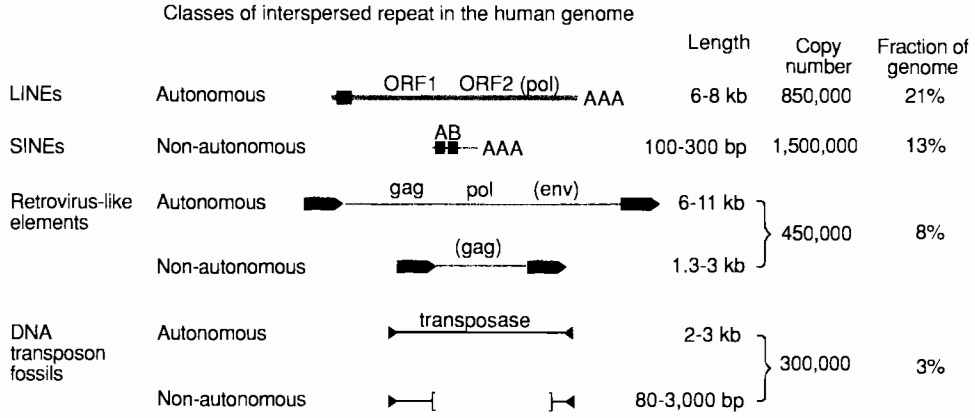


FIGURE 17.18. *There are four types of transposable elements in the human genome: LINEs, SINEs, LTR transposons, and DNA transposons. (From IHGSC, 2001.) Used with permission.*

(Jurka, 1998; Ostertag and Kazazian, 2001):

- LINEs occupy 21% of the human genome.
- SINEs occupy 13% of the human genome.
- LTR transposons account for 8% of the human genome.
- DNA transposons comprise about 3% of the human genome.

The structure of these repeats is shown in Figure 17.18, as well as their abundance in the human genome. LINEs, SINEs, and LTR transposons are all retrotransposons that encode a reverse transcriptase activity. They integrate into the genome through an RNA intermediate. In contrast, DNA transposons have inverted terminal repeats and encode a bacterial transposon-like transposase activity.

The number of interspersed repeats was estimated using RepeatMasker to search RepBase (see Chapter 16).

Retrotransposons can further be classified into those that are autonomous (encoding activities necessary for their mobility) and those that nonautonomous (depending on exogenous activities such as DNA repair enzymes). The most common nonautonomous retrotransposons are *Alu* elements.

Alu elements are so named because the restriction enzyme *Alu* I digests them in the middle of the sequence. In mouse, these are called B1 elements.

Interspersed repeats occupy a far greater proportion of the human genome than in other eukaryotic genomes (Table 17.10). The total number of interspersed repeats is estimated to be 3 million. These repeats offer an important opportunity to study molecular evolution. Each repeat element, even if functionally inactive, represents a "fossil record" which can be used to study genome changes within and between species. Transposons accumulate mutations randomly and independently.

TABLE 17-10 Interspersed Repeats in Four Eukaryotic Genomes

"Bases" refers to percentage of bases in the genome, "families" to approximate number of families in the genome.

	Human		*Drosophila*		*C. elegans*		*A. thaliana*	
	Bases	Families	Bases	Families	Bases	Families	Bases	Families
LINE/SINE	33.4%	6	0.7%	20	0.4%	10	0.5%	10
LTR	8.1%	100	1.5%	50	0%	4	4.8%	70
DNA	2.8%	60	0.7%	20	5.3%	80	5.1%	80
Total	44.4%	170	3.1%	90	6.5%	90	10.5%	160

Source: Adapted from IHGSC (2001). Used with permission.

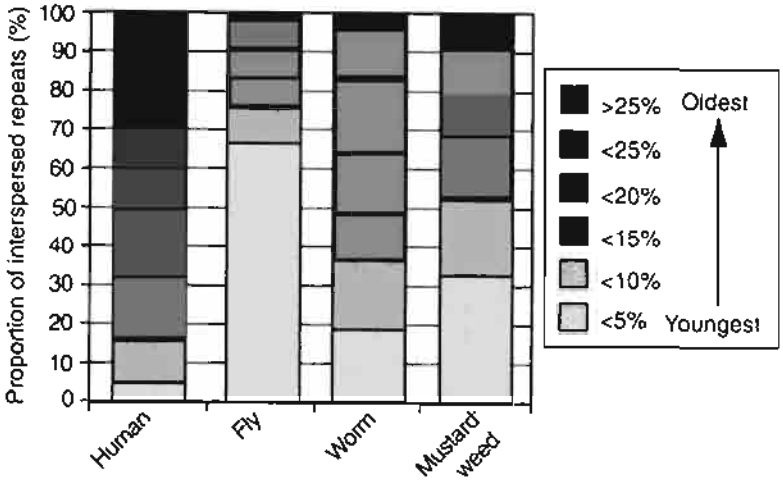


FIGURE 17.19. *Comparison of the age of interspersed repeats in four eukaryotic genomes. Humans have a small proportion of recent interspersed repeats. (From IHGSC, 2001.) Used with permission.*

It is possible to perform a multiple sequence alignment of transposons and to calculate the percent sequence divergence. Transposon evolution is assumed to behave like a molecular clock, which can be calibrated based on the known age of divergence of species such as humans and Old World monkeys [23 million years ago (MYA)]. Based on such phylogenetic analyses, several conclusions can be made (IHGSC, 2001) (Fig. 17.19):

- Most interspersed repeats in the human genome are ancient, predating the mammalian eutherian radiation 100 MYA. These elements are removed from the genome only slowly.
- SINEs and LINEs have long lineages, some dating back 150 MYA.
- There is no evidence for DNA transposon activity in the human genome in the past 50 million years; thus, they are extinct fossils.

Simple Sequence Repeats

Simple sequence repeats are repetitive DNA elements that consist of a perfect (or slightly imperfect) tandem repeats of k-mers. When the repeat unit is short (k is about one to a dozen bases), the simple sequence repeat is called a microsatellite. When the repeat unit is longer (from about a dozen to 500 bases), it is called a minisatellite (Toth et al., 2000).

Micro- and minisatellites comprise about 3% of the human genome (IHGSC, 2001). The most common repeat lengths are shown in Table 17.11. The most common repeat units are the dinucleotides AC, AT, and AG. We saw examples of these with the RepeatMasker program (Chapter 16).

Segmental Duplications

Segmental duplications occur when the human genome contains duplicated blocks of from 1 to 200 kb of genomic sequence. About 3.6% of the finished human genome sequence consists of segmental duplications, typically of 10–50 kb (Bailey et al., 2001). Many of these duplication events are recent, because both introns and coding regions are highly conserved. (For ancient duplication events, less conservation

TABLE 17-11 Simple Sequence Repeats (Microsatellites) in Human Genome

Length of Repeat	Average Bases per Megabase	Average Number of SSR Elements per Megabase
1	1660	33.7
2	5046	43.1
3	1013	11.8
4	3383	32.5
5	2686	17.6
6	1376	15.2
7	906	8.4
8	1139	11.1
9	900	8.6
10	1576	8.6
11	770	8.7

Source: IHGSC, 2001.
Abbreviation: SSR, simple sequence repeat.

is expected between duplicated intronic regions.) Segmental duplications may be interchromosomal or intrachromosomal. The centromeres contain large amounts of interchromosomal duplicated segments, with almost 90% of a 1.5-Mb region containing these repeats (Fig. 17.20). Smaller regions of these repeats also occur near the telomeres.

Gene Content of Human Genome

It is of great interest to characterize the gene content of the human genome because of the critical role of genes in human biology. However, the genes are the hardest

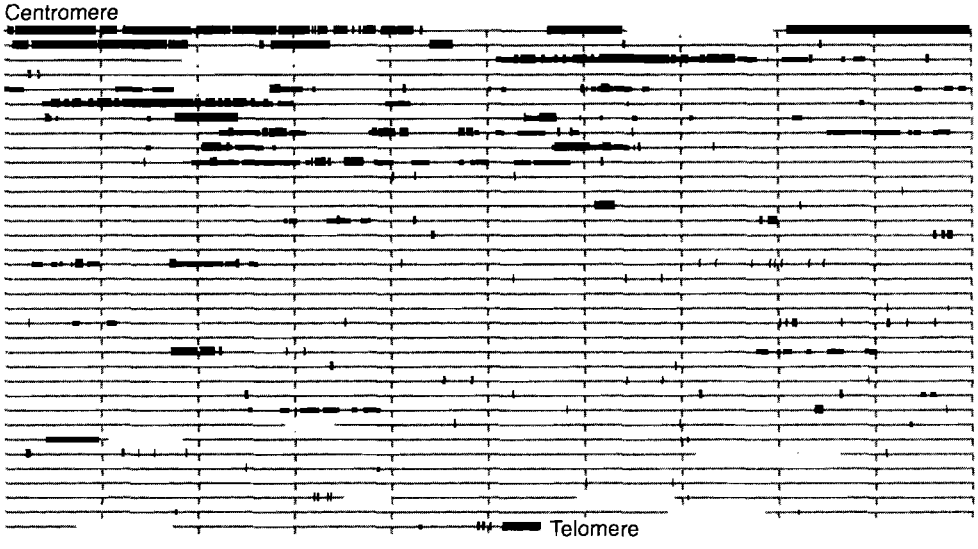


FIGURE 17.20. *The centromeres consist of large amounts of interchromosomal duplicated segments. The size and location of intrachromosomal (black) and interchromosomal (red) segmental duplications are indicated. Each horizontal line represents 1 Mb of chromosome 22q; the tick marks indicate 100-kb intervals. The centromere is at top left, and the telomere is at the lower right. (Adapted from IHGSC, 2001.) Used with permission.*

TABLE 17-12 Noncoding Genes in Human Genome

RNA Gene	Number of Noncoding Genes	Number of Related Genes	Function
tRNA	497	324	Protein synthesis
SSU (18S) RNA	0	40	Protein synthesis
5.8S rRNA	1	11	Protein synthesis
LSU (28S) rRNA	0	181	Protein synthesis
5S RNA	4	520	Protein synthesis
U1	16	134	Spliceosome component
U2	6	94	Spliceosome component
U4	4	87	Spliceosome component
U4atac	1	20	Minor (U11/U12) spliceosome component
U5	1	31	Spliceosome component
U6	44	1135	Spliceosome component
U6atac	4	32	Minor (U11/U12) spliceosome component
U7	1	3	Histone mRNA 3′ processing
U11	0	6	Minor (U11/U12) spliceosome component
U12	1	0	Minor (U11/U12) spliceosome component
SRP (7SL) RNA	3	773	Component of signal recognition particle
RNAse P	1	2	tRNA 5′ end processing
RNAse MRP	1	6	rRNA processing
Telomerase RNA	1	4	Template for addition of telomeres
hY1	1	353	Component of Ro RNP, function unknown
hY3	25	414	Component of Ro RNP, function unknown
hY4	3	115	Component of Ro RNP, function unknown
hY5 (4.5S RNA)	1	9	Component of Ro RNP, function unknown
Vault RNAs	3	1	Component of 13 Mda vault RNP
7SK	1	330	Unknown
H19	1	2	Unknown
Xist	1	0	Initiation of X chromosome inactivation
Known C/D snoRNAs	69	558	Pre-rRNA processing or site-specific ribose methylation of rRNA
Known H/ACA snoRNAs	15	87	Pre-rRNA processing or site-specific pseudouridylation of rRNA

Source: Adapted from IHGSC (2001). Used with permission.

features of genomic DNA to identify (see Chapter 16). This is a challenging task for many reasons:

- The average exon is only 50 codons (150 nucleotides). Such small elements are hard to identify as exons unambiguously.
- Exons are interrupted by introns, some many kilobases in length. In the extreme case, the human dystrophin gene extends over 2.4 Mb, the size of an entire genome of a typical prokaryote. Thus, the use of complementary DNAs continues to provide an essential approach to gene identification.
- There are many pseudogenes that may be difficult to distinguish from functional protein-coding genes.
- The nature of noncoding genes is poorly understood (see Chapter 16 and below).

Annemarie Poustka and colleagues recently sequenced 500 novel human cDNAs containing complete reading frames. They sequenced the 5′ end of over 30,000 cDNA clones (Wiemann et al., 2001). You can view the data via ▸ http://www.rzpd.de/.

Noncoding RNAs

There are many classes of human genes that do not encode proteins. Noncoding RNAs can be difficult to identify in genomic DNA because they lack open reding frames, they may be small, and they are not polyadenylated. Thus they are difficult to detect by gene-finding algorithms, and they are not present in cDNA libraries. These noncoding RNAs include the following:

- Transfer RNAs, required as adapters to translate mRNA into the amino acid sequence of proteins
- Ribosomal RNAs, required for mRNA translation
- Small nucleolar RNAs (snoRNAs), required for RNA processing in the nucleolus
- Small nuclear RNAs (snRNAs), required for spliceosome function

Hundreds of noncoding RNAs were identified in the draft version of the human genome (Table 17.12). The tRNA genes were most predominant, with 497 such genes and an additional 324 tRNA-derived pseudogenes. The tRNA genes associated with the human genetic code can now be described. This version of the genetic code includes the frequency of codon utilization for each amino acid and the number of tRNA genes that are associated with each codon. The total number of tRNA genes is comparable to that observed in other eukaryotes (Table 17.13).

TABLE 17-13 Estimated Number of tRNA Genes in Various Organisms

Organism	Canonical tRNAs	SeCys tRNA
Homo sapiens	497	1
Caenorhabditis elegans	584	1
Drosophila melanogaster	284	1
Saccharomyces cerevisiae	273	0
Methanococcus jannaschii	36	1
Escherichia coli	86	1

Abbreviation: SeCys, selenocysteine (see Box 3.4, page 54).

TABLE 17-14 Characteristics of Human Genes

Feature	Size (median)	Size (mean)
Internal exon	122 bp	145 bp
Exon number	7	8.8
Introns	1023 bp	3365 bp
3′ Untranslated region	400 bp	770 bp
5′ Untranslated region	240 bp	300 bp
Coding sequence	1100 bp	1340 bp
Coding sequence	367 aa	447 aa
Genomic extent	14 kb	27 kb

Source: Adapted from IHGSC (2001).
Abbreviation: aa, amino acids.

Protein-Coding Genes

The longest coding sequence is titin (80,780 bp; NM_003319). The gene for titin, on chromosome 2q24.3, has 178 exons and encodes a muscle protein of 26,926 amino acids (about 3 million Da). By contrast, a typical protein encoded by an mRNA of 1340 bp is about 50,000 Da.

Protein-coding genes are characterized by exons, introns, and regulatory elements. These basic features are summarized in Table 17.14. The average coding sequence for human genes is 1340 bp (IHGSC, 2001). This is comparable to the size of an average coding sequence in nematode (1311 bp) and *Drosophila* (1497 bp). Most internal exons are about 50–200 bp in length in all three species (Fig. 17.21*a*), although worm and fly have a greater proportion of longer exons (note the flatter tail in Fig. 17.21*a*). However, the size of human introns is far more variable (Figs. 17.21*b*, *c*). This results in a more variable overall gene size in humans than in worm and fly.

Chromosome 19, the most GC-rich chromosome, also houses the greatest density of genes (26.8 per megabase). The average density of gene predictions across the genome is 11.1 per megabase. The Y chromosome is least dense, having 6.4 predicted genes per megabase.

Protein-coding genes are associated with a high GC content (Fig. 17.22). While the overall GC content of the human genome is about 41%, the GC content of known genes (having RefSeq identifiers) is higher (Fig. 17.22*a*). Gene density increases 10-fold as GC content rises from 30 to 50% (Fig. 17.22*b*).

In an effort to catalog all protein-coding genes and their protein products, the IHGSC (2001) created an integrated gene index (IGI) and a corresponding integrated protein index (IPI). According to the European Bioinformatics Institute proteome analysis service, there are 33,817 proteins in the human proteome. Ensembl predictions are shown for the 15 most common protein families (Table 17.15), most common domains (Table 17.16), and most common repeats (Table 17.17).

EBI proteome analysis is available at ▸ http://www.ebi.ac.uk/proteome/.

Comparative Proteome Analysis

The importance of comparative analyses has emerged as one of the fundamental tenets of genomics. A comparison of human proteins to proteins from the completed genomes of *S. cerevisiae, A. thaliana, C. elegans,* and *D. melanogaster* is shown in Table 17.18. The IHGSC (2001) analyzed functional groups of these proteins based on InterPro and Gene Ontology (GO) Consortium classifications. Humans have relatively more genes that encode proteins predicted to function in cytoskeleton, transcription/translation, and defense and immunity (Fig. 17.23).

We discussed GO Consortium and InterPro in Chapters 8 and 10.

The human proteome was further studied by blastp searching every predicted protein against the nonredundant database. The distribution of homologs is shown in Figure 17.24. Overall, 74% of the proteins were significantly related to other known proteins. As more sequences are accumulated in databases over time, the matches between human proteins and other eukaryotes (and prokaryotes) will continue to increase.

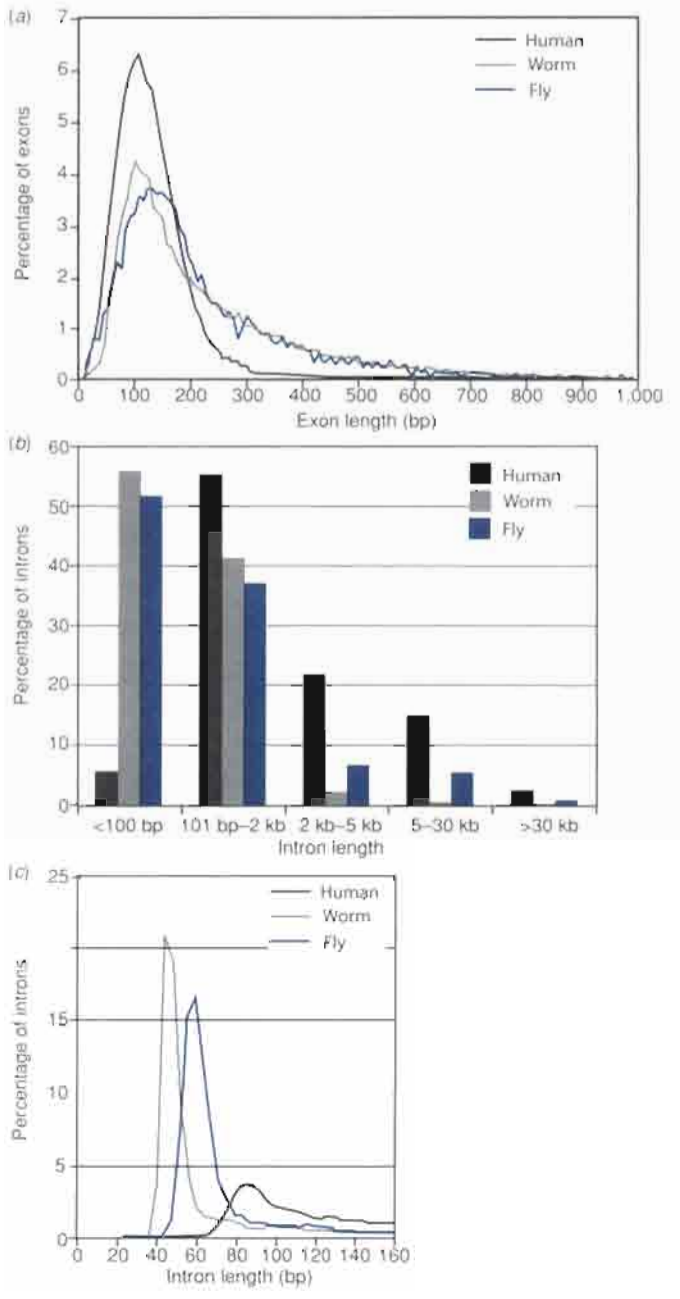


FIGURE 17.21. *Size distribution of* (a) *exons,* (b) *introns, and* (c) *short introns [enlarged from* (b)*] in human, worm, and fly. (From IHGSC, 2001.) Used with permission.*

Complexity of Human Proteome

The number of protein-coding genes in humans is comparable to the number of genes in other metazoans and plants and only fivefold greater than the number in unicellular fungi. Nonetheless the human proteome may be far more complex for several reasons (IHGSC, 2001):

1. There are relatively more domains and protein families in humans than in other organisms.

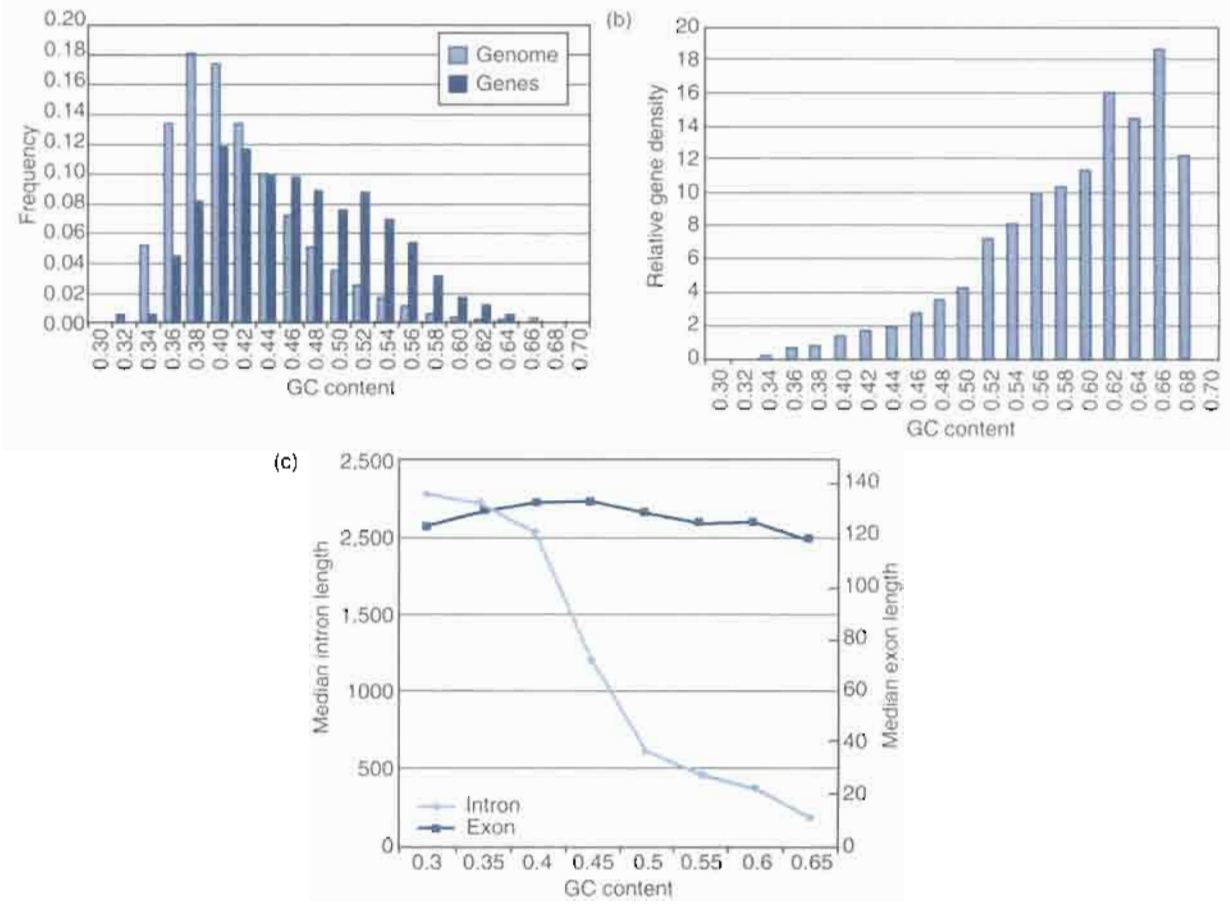


FIGURE 17.22. (a) *Distribution of GC content in genes and in the genome shows that protein-coding genes are associated with a higher GC content.* (b) *The gene density is plotted as a function of the GC content. [The density is obtained by taking the ratio of the values in* (a).*] As GC content rises, the relative gene density increases dramatically.* (c) *Mean exon length is unaffected by GC content, but introns are far shorter as GC content rises. (From IHGSC, 2001.) Used with permission.*

TABLE 17-15 Fifteen Most Common Families for *Homo sapiens*

InterPro	Proteins Matched	Name
IPR000276	837	Rhodopsin-like GPCR superfamily
IPR000719	685	Protein kinase
IPR001909	305	KRAB box
IPR001806	191	Ras GTPase superfamily
IPR005821	153	Ion transport protein
IPR000387	129	Tyrosine-specific protein phosphatase and dual-specificity protein phosphatase
IPR001254	126	Serine protease, trypsin family
IPR000379	107	Esterase/lipase/thioesterase, active site
IPR001664	83	Intermediate filament protein
IPR001128	79	Cytochrome P450
IPR000910	78	HMG1/2 (high-mobility group) box
IPR000832	73	G-protein coupled receptors family 2 (secretin-like)
IPR002198	72	Short-chain dehydrogenase/reductase (SDR)
IPR005828	72	General substrate transporter
IPR001394	67	Ubiquitin thiolesterase, family 2

Source: From ▸http://www.ebi.ac.uk/proteome/, November 2002.

TABLE 17-16 Fifteen Most Common Domains for *Homo sapiens*

InterPro	Matches per Genome[a]	Name
IPR000822	27,858 (912)	Zinc finger, C2H2 type
IPR003006	2,329 (853)	Immunoglobulin/major histocompatibility complex
IPR002965	1,525 (364)	Proline-rich extensin
IPR001841	1,090 (361)	Zinc finger, RING
IPR001849	1,054 (330)	Pleckstrin-like
IPR000504	1,620 (317)	RNA-binding region RNP-1 (RNA recognition motif)
IPR006209	2,942 (299)	EGF-like domain
IPR001452	1,814 (282)	SH3 domain
IPR002048	1,682 (263)	Calcium-binding EF-hand
IPR001356	1,403 (251)	Homeobox
IPR003961	1,819 (229)	Fibronectin, type III
IPR001478	1,097 (208)	PDZ/DHR/GLGF domain
IPR000210	498 (194)	BTB/POZ domain
IPR005225	188 (188)	Small GTP-binding protein domain
IPR002126	4,168 (169)	Cadherin

Source: From ▶http://www.ebi.ac.uk/proteome/, November 2002.
[a] Numbers in parentheses are proteins matched.

TABLE 17-17 Fifteen Most Common Repeats for *Homo sapiens*

InterPro	Proteins Matched	Name
IPR001680	364	G-protein beta WD-40 repeat
IPR002110	263	Ankyrin
IPR001611	215	Leucine-rich repeat
IPR001440	173	TPR repeat
IPR000087	97	Collagen triple-helix repeat
IPR001798	95	Kelch repeat
IPR002017	80	Spectrin repeat
IPR000884	76	Thrombospondin, type I
IPR003659	59	Plexin/semaphorin/integrin
IPR002172	53	Low-density lipoprotein receptor, class A domain
IPR000225	46	Armadillo repeat
IPR001610	34	PAC motif
IPR002165	30	Plexin
IPR000357	28	HEAT repeat
IPR000585	25	Hemopexin repeat

Source: From ▶http://www.ebi.ac.uk/proteome/, November 2002.

TABLE 17-18 Proteome Comparisons between Human and *Arabidopsis, C. elegans, Drosophila,* and *S. cerevisiae*

Organism	Number of Proteins in Proteome	Proteins with InterPro Matches	Percent of all Proteins	Number of Signatures	Number of InterPro Entries
H. sapiens	28,469	20,530	72.1	5,993	3,193
M. musculus	21,001	15,898	75.7	5,733	3,046
A. thaliana	18,064	13,684	75.8	4,103	2,168
C. elegans	20,506	14,421	70.3	4,316	2,251
D. melanogaster	13,774	10,460	75.9	4,409	2,279
S. cerevisiae	6,159	4,192	68.1	3,578	1,816

Source: From ▸ http://www.ebi.ac.uk/proteome/, November 2002.

2. The human genome encodes relatively more paralogs, potentially yielding more functional diversity.
3. There are relatively more multidomain proteins having multiple functions.
4. Domain architectures tend to be more complex in the human proteome.
5. Alternative RNA splicing may be more extensive in humans.

There may be a synergistic effect among these factors, leading to a substantially greater complexity of the human proteome that could account for the phenotypic complexity of vertebrates, including humans.

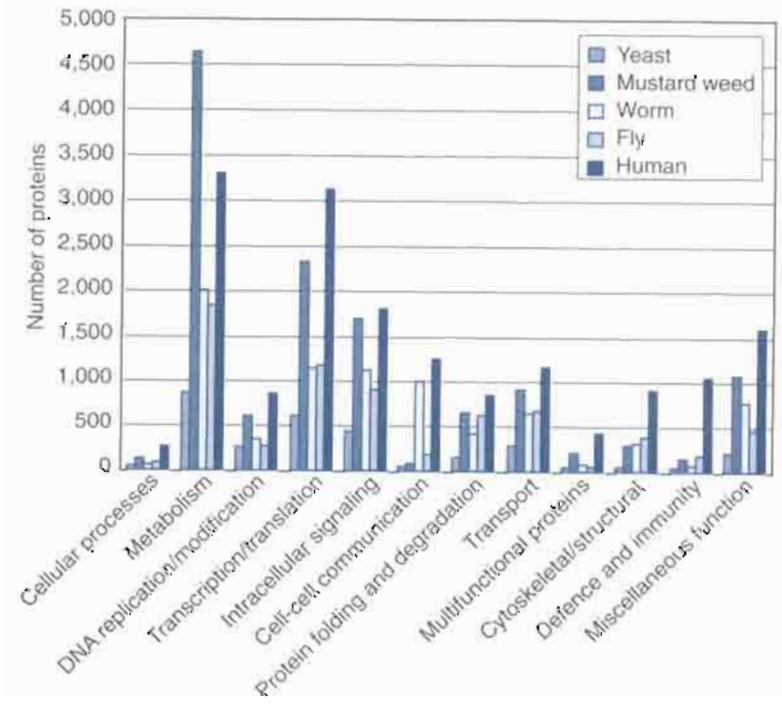


FIGURE 17.23. *Functional categories in eukaryotic proteomes of yeast* (S. cerevisiae), *mustard weed* (A. thaliana), *worm* (C. elegans), *Fly* (D. melanogaster), *and human. The classification categories were derived from InterPro (Chapter 10). (From IHGSC, 2001.) Used with permission.*

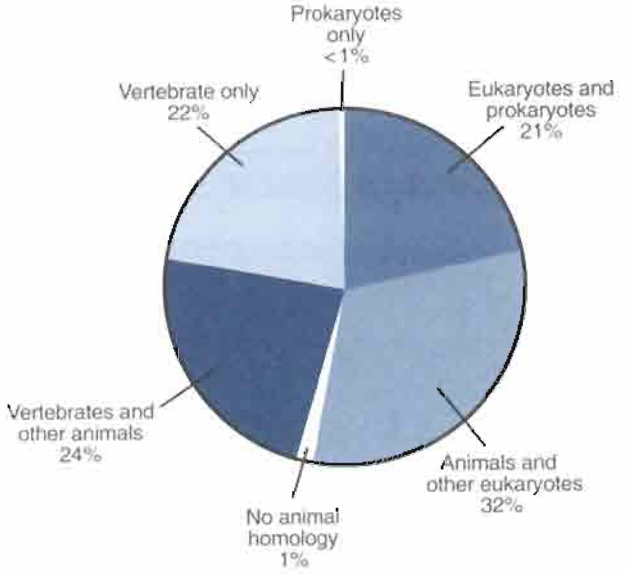


FIGURE 17.24. *Taxonomic distribution of the protein homologs of predicted human proteins. Each protein was searched by blastp, and proteins with an* E *value less than 0.001 were called homologs. Additional PSI-BLAST searches were performed (with three iterations). (From IHGSC, 2001.) Used with permission.*

PERSPECTIVE

The sequencing of the human genome represents one of the great accomplishments in the history of science. This effort is the culmination of decades of work in an international effort. Two major technological advances enabled the human genome to be sequenced: (1) the invention of automated DNA sequencing machines in the 1980s allowed nucleotide data to be collected on a large scale and (2) the computational biology tools necessary to analyze those sequence data were created by biologists and computer scientists. In the coming years, we can expect the pace of DNA sequence to continue to increase. It may soon be possible to compare the complete genome sequence of many individuals in an effort to relate genotype to phenotype.

PITFALLS

There are a number of outstanding problems that have yet to be solved:

- How can we accurately determine the number of protein-coding genes?
- How can we determine the number of noncoding genes?
- How can we determine the function of genes and proteins?
- What is the evolutionary history of our species?
- What is the degree of heterogeneity between individuals at the nucleotide level?

Thus, as we take our first look at the human genome, it is appropriate to see this moment as a beginning rather than an end. Having the sequence in hand, and having the opportunity to compare the human genome sequence to that of many other genomes, we are now in a position to pose a new generation of questions.

WEB RESOURCES

TABLE 17-19 Additional Websites for Publicly Available Human Genome Sequence Data

Resource	Description	URL
National Center for Biotechnology Information	Views of chromosomes, maps, and loci	▸http://ncbi.nlm.nih.gov/genome/guide
Oak Ridge National Laboratory	Views of human genome	▸http://compbio.ornl.gov/channel/index.html
RIKEN and the University of Tokyo	Overview of human genome	▸http://hgrep.ims.u-tokyo.ac.jp/
Washington University	Links to clone and accession maps of human genome	▸http://genome.wustl.edu/gsc/human/Mapping/

DISCUSSION QUESTIONS

[17-1] If you had the resources and facilities to sequence the entire genome of five individuals, which would you select? Why?

[17-2] The *Saccharomyces cerevisiae* genome duplicated about 100 MYA, as indicated by BLAST searching (Chapter 15). Why is it not equally straightforward to identify large duplications of the human genome? Is it because they did not occur, or because the evolutionary history of humans obscures such events, or because we lack the tools to detect such large-scale genomic changes? For a thoughtful discussion of duplications in the human genome, see an article by Evan Eichler (2001).

PROBLEM

[17-1] Go to LocusLink, and select a human gene of interest, or else type "lipocalin" and choose a gene. Examine the features of this gene at the Ensembl, NCBI, and UCSC websites. Make a table of various properties (e.g. exon/intron structure, number of ESTs corresponding to the expressed gene, polymorphisms identified in the gene, neighboring genes). Are there discrepancies between the data reported in the three databases? Next, obtain a portion of genomic DNA (about 100,000 base pairs in the FASTA format) from the region including this gene. Use the Oak Ridge National Laboratory pipeline to characterize the genomic DNA and potential protein-coding regions as described in Figures 16.13 to 16.16. The URL is ▸http://compbio.ornl.gov/tools/pipeline.

SELF-TEST QUIZ

[17-1] Approximately how large is the human genome?
(a) 130 Mb
(b) 300 Mb
(c) 3000 Mb
(d) 30,000 Mb

[17-2] Approximately what percentage of the human genome consists of repetitive elements of various kinds?
(a) 5%
(b) 25%
(c) 50%
(d) 95%

[17-3] What percentage of the human genome is devoted to the protein-coding regions?
(a) 1–5%
(b) 5–10%
(c) 10–20%
(d) 20–40%

[17-4] The UCSC human genome browser differs from the Ensembl and NCBI human genome sites because
(a) It offers a large number of annotation tracks, about half of which are supplied by external users of the site
(b) It offers a large number of chromosome maps, including maps of conserved syntenic regions
(c) It offers a genome assembly based on BLAST
(d) It offers a genome assembly incorporating both public consortium data and Celera data

[17-5] The human genome contains many transposon-derived repeats. These are described as
(a) dead fossils
(b) young, active elements

(c) human-specific elements
(d) inverted repeats

[17-6] Approximately how much of the human genome do segmental duplications occupy?
(a) <1%
(b) 3–5%
(c) 20–30%
(d) 50%

[17-7] In areas of high GC content of the human genome,
(a) gene density tends to be low
(b) gene density tends to be high
(c) gene density is variable
(d) genes tend to have fewer introns

[17-8] In comparison to other metazoan genomes (such as nematodes, insects and mouse),
(a) the human genome contains considerably more protein-coding genes
(b) the human genome has considerably more unique genes that lack identifiable orthologs
(c) the human genome has a higher GC content
(d) the human genome has somewhat more multidomain proteins, paralogous genes, and alternative splicing.

SUGGESTED READING

In this chapter, we focused on the public consortium description of the human genome (IHGSC, 2001). The companion Celera article (Venter et al., 2001) is also of great interest, as are the many accompanying articles in those issues of *Science* and *Nature*.

The first three human chromosomes to be completely sequenced were chromosomes 22 (Dunham et al., 1999), 21 (Hattori et al., 2000), and 20 (Deloukas et al., 2001). These important papers describe the in-depth analyses of finished (or nearly finished) chromosomal sequences. They highlight the need for complete sequencing in order to perform more accurate annotation and comparative analyses.

There is of course a large body of literature on the human genome. The May 2001 issue of *Genome Research* contains articles by Francis Collins, Eric Green, and Aravinda Chakravarti providing important perspectives on the current status of the Human Genome Project.

A good starting point to read about the Human Genome Project is at the National Human Genome Institute website, ▸http://genome.gov.

REFERENCES

Amir, R. E., et al. Rett syndrome is caused by mutations in X-linked *MECP2,* encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23,** 185–188 (1999).

Avner, P., and Heard, E. X-chromosome inactivation: Counting, choice and initiation. *Nat. Rev. Genet.* **2,** 59–67 (2001).

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J., and Eichler, E. E. Segmental duplications: Organization and impact within the current human genome project assembly. *Genome Res.* **11,** 1005–1017 (2001).

Baltimore, D. Our genome unveiled. *Nature* **409,** 814–816 (2001).

Bernardi, G. Misunderstandings about isochores. Part 1. *Gene* **276,** 3–13 (2001).

Crick, F. H. and Watson, J. D. Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (1953).

Deloukas, P., et al. The DNA sequence and comparative analysis of human chromosome 20. *Nature* **414,** 865–871 (2001).

Dunham, I., et al. The DNA sequence of human chromosome 22. *Nature* **402,** 489–495 (1999).

Eichler, E. E. Segmental duplications: What's missing, misassigned, and misassembled–and should we care? *Genome Res.* **11,** 653–656 (2001).

Gardiner-Garden, M., and Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196,** 261–282 (1987).

Goodman, N. The IT GUY. *Genome Technol.* **1,** 55–59 (2002).

Green, E. D. Strategies for the systematic sequencing of complex genomes. *Natl. Rev. Genet.* **2,** 573–583 (2001).

Green, E. D., and Chakravarti, A. The human genome sequence expedition: Views from the "base camp." *Genome Res.* **11,** 645–651 (2001).

Haring, D., and Kypr, J. Mosaic structure of the DNA molecules of the human chromosomes 21 and 22. *Mol. Biol. Rep.* **28,** 9–17 (2001).

Hattori, M., et al. The DNA sequence of human chromosome 21. The chromosome 21 mapping and sequencing consortium. *Nature* **405,** 311–319 (2000).

Hubbard, T., et al. The Ensembl genome database project. *Nucleic Acids Res.* **30,** 38–41 (2002).

IHGSC. Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

Jurka, J. Repeats in genomic DNA: Mining and meaning. *Curr. Opin. Struct. Biol.* **8,** 333–337 (1998).

Kent, W. J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12,** 656–664 (2002).

Kent, W. J., et al. The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

Li, S., et al. Comparative analysis of human genome assemblies reveals genome-level differences. *Genomics* **80,** 138–139 (2002).

National Research Council. *Mapping and Sequencing the Human Genome.* National Academy Press, Washington, DC, 1988.

Ostertag, E. M., and Kazazian, H. H., Jr. Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* **11,** 2059–2065 (2001).

Pavlicek, A., Paces, J., Clay, O., and Bernardi, G. A compact view of isochores in the draft human genome sequence. *FEBS Lett.* **511,** 165–169 (2002).

Ponting, C. P. Plagiarized bacterial genes in the human book of life. *Trends Genet.* **17,** 235–237 (2001).

Ross, D. T., et al. Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* **24,** 227–235 (2000).

Salzberg, S. L., White, O., Peterson, J., and Eisen, J. A. Microbial genes in the human genome: Lateral transfer or gene loss? *Science* **292,** 1903–1906 (2001).

Singer, C. *The Fasciculo Medicina Venice 1493.* R. Lier and Co., Florence, 1925.

Toth, G., Gaspari, Z., and Jurka, J. Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res.* **10,** 967–981 (2000).

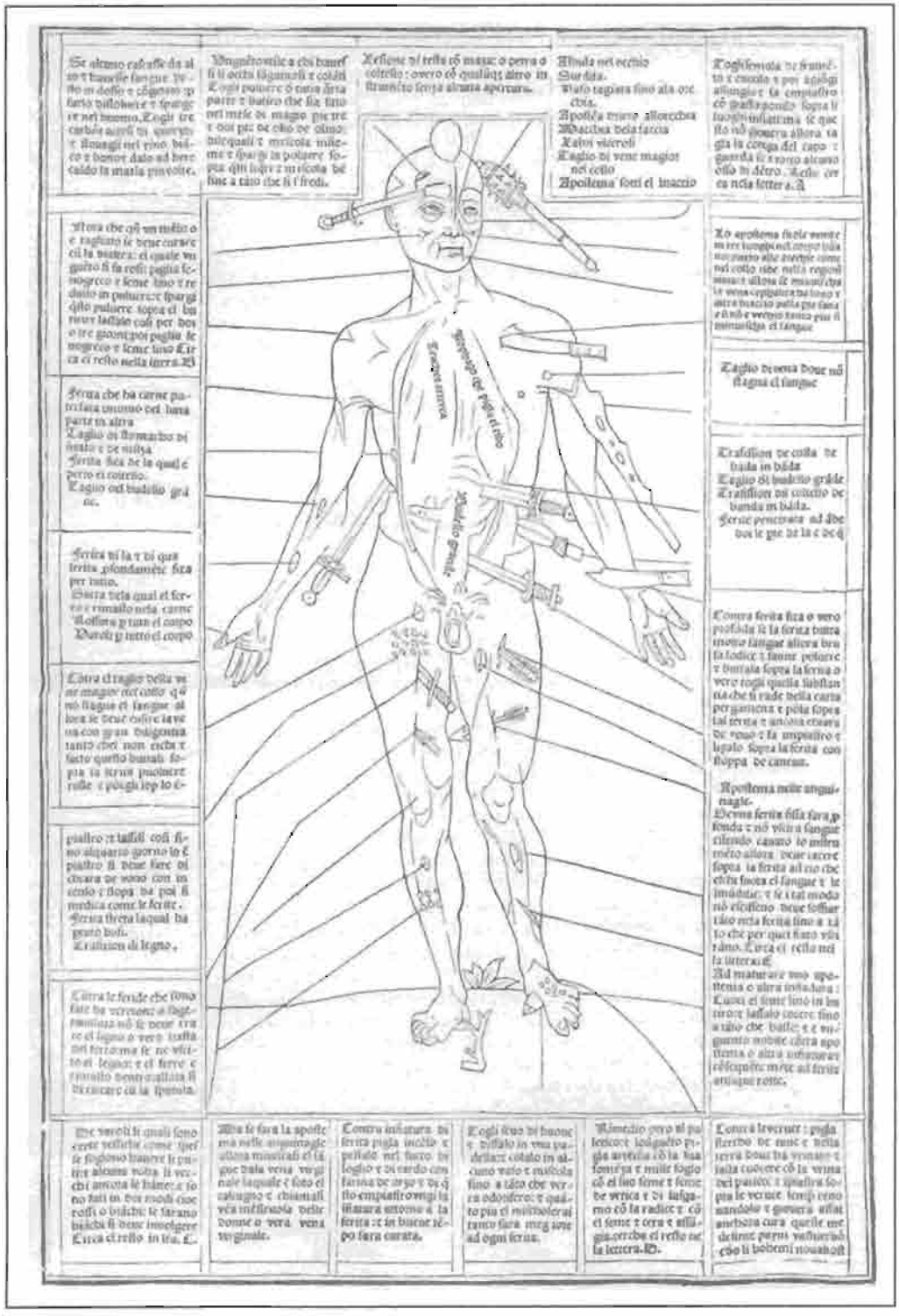
Tycko, B., and Morison, I. M. Physiological functions of imprinted genes. *J. Cell. Physiol.* **192,** 245–258 (2002).

Venter, J. C., et al. The sequence of the human genome. *Science* **291,** 1304–1351 (2001).

Wiemann, S., et al. Toward a catalog of human genes and proteins: Sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.* **11,** 422–435 (2001).

Yu, A., et al. Comparison of human genetic and sequence-based physical maps. *Nature* **409,** 951–953 (2001).

This Page Intentionally Left Blank



The Fasciculus medicinae *(first printed 1491), attributed to Johannes de Ketham, includes a "wound man." This figure provides procedures and recipes for a variety of ills. The most common medical procedure of the time was probably venesection. According to the text at bottom left: "On varioles, as are called those little blisters which children often tend to get, and which older people sometimes have: these occur in two ways, some are white while others are red. If they are white the entire body must be wrapped in a linen cloth." At top left the text reads: "For a fall from high, as when someone has fallen from a high place and his blood has coagulated and then this blood is spread and dissolved in man: take three live oak charcoals, extinguish them in good white wine, and let the patient repeatedly drink this in the morning." See Singer (1925).*

18

Human Disease

> Life is a relationship between molecules, not a property of any one molecule. So is therefore disease, which endangers life. While there are molecular diseases, there are no diseased molecules. At the level of the molecules we find only variations in structure and physicochemical properties. Likewise, at that level we rarely detect any criterion by virtue of which to place a given molecule "higher" or "lower" on the evolutionary scale. Human hemoglobin, although different to some extent from that of the horse (Braunitzer and Matsuda, 1961), appears in no way more highly organized. Molecular disease and evolution are realities belonging to superior levels of biological integration. There they are found to be closely linked, with no sharp borderline between them. The mechanism of molecular disease represents one element of the mechanism of evolution. Even subjectively the two phenomena of disease and evolution may at times lead to identical experiences. The appearance of the concept of good and evil, interpreted by man as his painful expulsion from Paradise, was probably a molecular disease that turned out to be evolution. Subjectively, to evolve must most often have amounted to suffering from a disease. And these diseases were of course molecular.
>
> —Emile Zuckerkandl and Linus Pauling (1962, pages 189–190)

Human Genetic Disease: A Consequence of DNA Variation

Variation in DNA sequence is a defining feature of life on Earth. For each species, genetic variation is responsible for the adaptive changes that underlie evolution. Evolution is a process by which species adapt to their environment. When changes in DNA improve the fitness of a species, its population reproduces more successfully. When changes are relatively maladaptive, the species may become extinct. At the level of the individual within a species, some mutations improve fitness, most

Bioinformatics and Functional Genomics by Jonathan Pevsner
ISBN 0-471-21004-8

TABLE 18-1 Mechanisms of Genetic Mutation. AG/GT Indicates Mutations in the Canonical First Two and Last Two Base Pairs of an Intron. Outside AG/GT Indicates Mutations in Less Canonical Sequences

Mechanism	Usual Effect	Example
Large Mutation		
Deletion	Null	Duchenne dystrophy
Insertion	Null	Hemophila A/LINE
Duplication	Null, gene disrupted	Duchenne dystrophy
	Dosage, gene intact	Charcot-Marie-Tooth
Inversion	Null	Hemophila A
Expanding triplet	Null	Fragile X
	Gain of function?	Huntington
Point Mutation		
Silent	None	Cystic fibrosis
Missense or in-frame deletion	Null, hypomorphic, altered function, benign	Globin
Nonsense	Null	Cystic fibrosis
Frame shift	Null	Cystic fibrosis
Splicing (AG/GT)	Null	Globin
Splicing (outside AG/GT)	Hypomorphic	Globin
Regulatory (TATA, other)	Hypomorphic	Globin
Regulatory (poly A site)	Hypomorphic	Globin

Source: Adapted from Beaudet et al. (2001. p. 9). Used with permission.

Mutation is the alteration of DNA sequence. The cause may be errors in DNA replication or repair, the effects of chemical mutagens, or radiation. While there may be negative connotations associated with the concept of mutations, mutation and fixation are the essential driving forces behind evolution.

From a medical perspective, disease is "a pathological condition of the body that presents a group of clinical signs, symptoms, and laboratory findings peculiar to it and setting the condition apart as an abnormal entity differing from other normal or pathological condition"(Thomas, 1993, p. 552). Disorder is a "pathological condition of the mind or body" (Thomas, 1993, p. 559). A syndrome is "a group of symptoms and signs of disordered function related to one another by means of some anatomical, physiological, or biochemical peculiarity. This definition does not include a precise cause of an illness but does provide a framework of reference for investigating it" (Thomas, 1993, p. 1185).

mutations have no effect on fitness, and some are maladaptive (relative to some norm). Disease may be defined as maladaptive changes that afflict individuals within a population. Disease is also defined as an abnormal condition in which physiological function is impaired. Our focus is on the molecular basis of physiological defects at the levels of DNA, RNA, and protein.

There is a tremendous diversity to the nature of human diseases. This is for several reasons:

- Mutations affect all parts of the human genome. There are limitless opportunities for maladaptive mutations to occur. These may be point mutations, affecting just a single nucleotide, or large mutations, affecting as much as an entire chromosome or multiple chromosomes.
- There are many mechanisms by which mutations can cause disease (summarized in Table 18.1). These include disruptions of gene function by point mutations that change the identity of amino acid residues; by deletions or insertions of DNA, ranging in size from one nucleotide to an entire chromosome that is over 100 million base pairs (Mb); or inversions of the orientation of a DNA fragment. In many cases, different kinds of mutations affecting the same gene cause distinct phenotypes.
- Most genes function by producing a protein as a gene product. A disease-causing mutation in a gene results in the failure to produce the gene product with normal function. This has profound consequences on the ability of the cells in which the gene product is normally expressed to function.
- The interaction of an individual with his or her environment has profound effects on disease phenotype. Genetically identical twins may have entirely different phenotypes. Such differences are attributable to environmental influences or to epigenetic effects. The concordance rate between monozygotic

twins is an indication of the relative extent to which genetic and environmental effects influence disease. Even for highly genetic disorders, such as autism (Box 18.1 below) and schizophrenia, the concordance rate is never 100%.

A Bioinformatics Perspective on Human Disease

In Chapter 1, we defined bioinformatics as a discipline that uses computer databases and computer algorithms to analyze proteins, genes, and genomes. Our approach to human disease is reductionist, in that we seek to describe genes and gene products that cause disease. However, an appreciation of the molecular basis of disease may be integrated with a holistic approach to uncover the logic of disease in the entire human population (Childs and Valle, 2000). As we explore bioinformatics approaches to human disease, we are constantly faced with the complexity of all biological systems. Even when we uncover the gene that when mutated causes a disease, our challenge is to attempt to connect the genotype to the phenotype. We can only accomplish this by synthesizing information about the biological context in which each gene functions and in which each gene product contributes to cellular function (Childs and Valle, 2000; Dipple et al., 2001).

The field of bioinformatics offers approaches to human disease that may help us to understand basic questions about the influence of genes and the environment on all aspects of the disease process. Some examples of ways in which this field can have an impact on our knowledge of disease will be highlighted throughout the chapter, and include the following.

- To the extent that the genetic basis of disease is a function of variation in DNA sequences, DNA databases offer us the basic material necessary to compare DNA sequences. These databases include major, general repositories of DNA sequence such as GenBank/EMBL/DDBJ (Chapter 2), general resources such as Online Mendelian Inheritance in Man (OMIM), and locus-specific databases that provide data on sequence variations at individual loci.
- Geneticists who search for disease-causing genes through linkage studies, association studies, or other tests (described below) depend on physical and genetic maps in their efforts to identify mutant genes.
- When a protein-coding gene is mutated, there is a consequence on the three-dimensional structure of the protein product. Bioinformatics tools described in Chapter 9 allow us to predict the structure of protein variants, and from such analyses we may infer changes in function.
- Once a mutant gene is identified, we want to understand the consequence of that mutation on cellular function. We have described a variety of approaches to understanding protein function in Chapter 8. And in our discussion of *Saccharomyces cerevisiae,* we discussed additional high throughput approaches to understanding eukaryotic protein function (Chapter 15). Gene expression studies (Chapters 6 and 7) have been employed to study the transcriptional response to disease states.
- We may obtain great insight into the role of a particular human gene by identifying orthologs in simpler organisms. We will discuss orthologs of human disease genes found in a variety of model systems.

In this chapter, we will first provide an overview of human disease, including approaches to disease classification. Next, we will consider the subject of human disease at three levels (outlined in Fig. 18.1). First, at the molecular level, we will focus

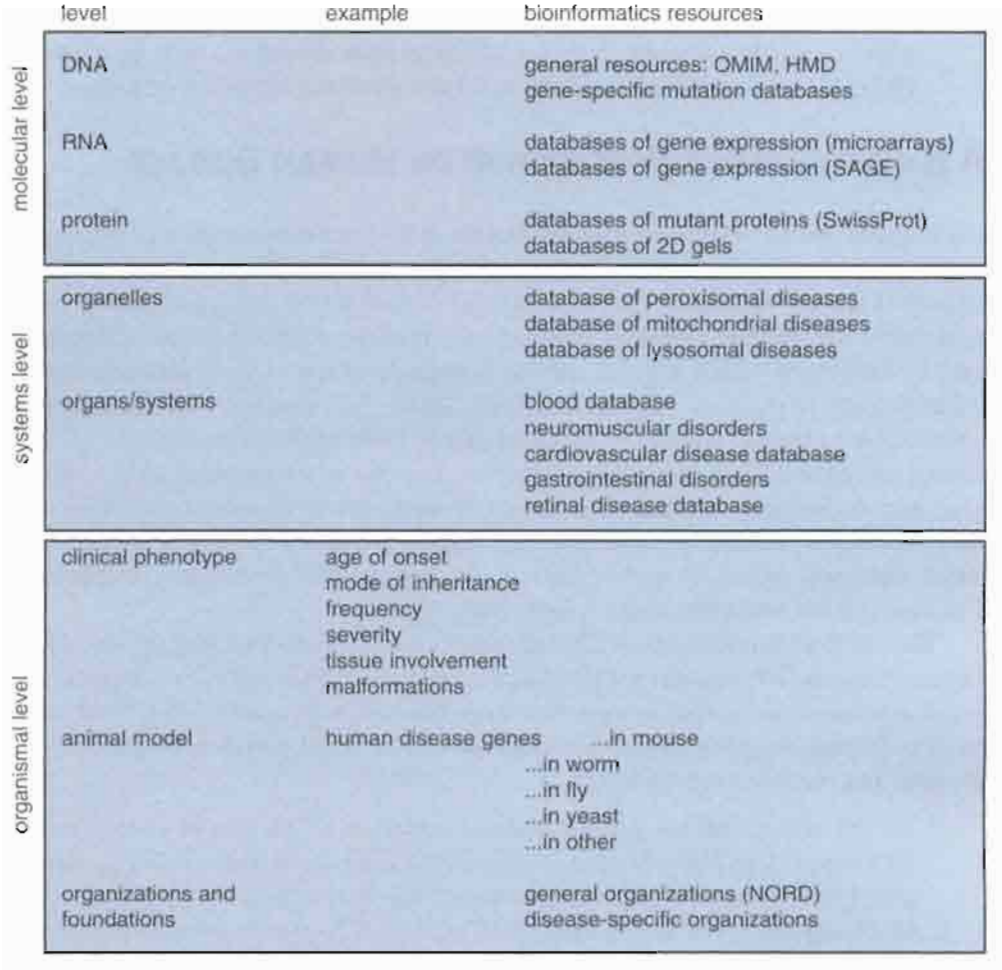


FIGURE 18.1. *Bioinformatics resources for the study of human disease are organized at a variety of levels.*

on the role of genes in disease. In discussing monogenic (single-gene) disorders, we will introduce Online Mendelian Inheritance in Man (OMIM), which is the principal disease database. There are also several hundred locus-specific mutation databases, and we will discuss these. We will also examine both bioinformatics approaches and databases relevant to the study of RNA, and protein. Second, we will examine web resources for diseases at the cellular and systems level, such as organellar disease databases. Third, we will consider the level of the organism: what bioinformatic tools have been developed to characterize the clinical phenotype of disease (e.g., age of onset, mode of inheritance, frequency, and severity)? What animal models of disease have been developed? We will explore orthologs of human disease genes in model organisms such as fungi and lower metazoans. Finally, we will consider databases that have been established to provide general information on human disease.

GARROD'S VIEW OF DISEASE

You can read Garrod's 1902 paper on alkaptonuria on-line at ▸http://www.ssiem.org.uk/garrod.pdf.

Sir Archibald Garrod (1857–1936) made important contributions to our understanding of the nature of human disease. In a 1902 paper, Garrod described his

studies of alkaptonuria, a rare inherited disorder. In alkaptonuria, the enzyme homogentisate 1,2-dioxygenase (HGD) is defective or missing. As a result, the amino acids phenylalanine and tyrosine cannot be metabolized properly, and a metabolite (homogentisic acid) accumulates. This metabolite oxides in urine and turns dark. Garrod considered this phenotype from the perspective of evolution, noting the influence of natural selection on chemical processes. Variations in metabolic processes between individuals might include those changes that cause disease.

The OMIM entry for alkaptonuria is #20355; the # sign is defined in Table 18.5 below. The RefSeq accession of HGD is NP_000178. The gene is localized to chromosome 3q21-q23.

Garrod had the insight that for each of the rare disorders he studied, the disease phenotype reflects the chemical individuality of the individual. He further realized that this trait was inherited—he proposed that alkaptonuria is transmitted by recessive Mendelian inheritance. At the time, it was thought that most diseases were caused by external forces such as bacterial infection. In studying this and related recessive disorders (such as cystinuria and albinism) he instead proposed that the manifestation of the disease is caused by an inherited enzyme deficiency or biochemical error (Scriver and Childs, 1989). He described this point of view in his first book, *Inborn errors of metabolism* (1909). Garrod wrote in 1923 (cited in Scriver and Childs, 1989, p. 7):

"If it may be granted that the individual members of a species vary from the normal of the species in chemical structure and chemical behaviour, it is obvious that such variations or mutations are capable of being perpetuated by natural selection; and not a few biologists of the present day assign to chemical structure and function a most important share in the evolution of species... Very few individuals exhibit such striking deviations from normal metabolism as porphyrinurics and cystinurics show, but I suspect strongly that minimal deviations which escape notice are almost universal. How else can be explained the part played by heredity in disease? There are some diseases which are handed down from generation to generation... which tend to develop in later childhood and early adult life... It is difficult to escape the conclusion that although these maladies are not congenital, their underlying causes are inborn peculiarities."

Garrod thus presented a new view of how inborn factors cause disease. He worked at a time before Beadle and Tatum offered the hypothesis that one gene encodes one protein, and Garrod never used the word "gene." But we now understand that the "inborn peculiarities" he described are mutated genes. A main conclusion of his work is that chemical individuality, achieved through genetic differences, is a major determinant of human health and disease. Although the phrase "chemical individuality" is not used often today, the concept is of tremendous interest in the field of pharmacogenomics. Not everyone who is exposed to an infectious agent gets sick, and it is imperative to understand why. Not everyone who takes a drug responds in a similar way.

Garrod further developed these ideas in a second book, *Inborn Factors in Disease* (1931). Here he addressed the question of why certain individuals are susceptible to diseases—whether the disease is clearly inherited or whether it derives from another cause such as an environmental agent. He argued that chemical individuality predisposes us to disease. Every disease process is affected by both internal and external forces: our genetic complement and the environmental factors we face. In some cases, such as inborn errors of metabolism, genetic factors have a more prominent role. In other cases, such as multifactorial disease, mutations in many genes are responsible for the disease. And in infectious disease, genes also have an important role in defining the individual's susceptibility and bodily response to the infectious agent. We will next proceed to discuss these various kinds of disease.

Disease category	Incidence/ thousand
single gene disorders (Mendelian)	11
autosomal dominant	6
autosomal recessive	3
X-linked recessive	1
X-linked mental retardation	1
complex (multigenic) disorders	630
congenital anomalies	30
central nervous system disorders	100
cardiovascular	500
chromosomal disorders	high
infectious disease	high
viral, bacterial, or other pathogen example: HIV-1/AIDS	
environmental disease	high
exogenous agent e.g. toxin, sunlight (non-genetic cause) example: lead poisoning	

FIGURE 18.2. *Human disease can be categorized based on the cause. These include single-gene disorders (mutations in a single gene; examples include phenylketonuria and sickle cell anemia); complex disorders (having mutations in two or more genes, such as cancer or schizophrenia); chromosomal disorders such as Down syndrome; infectious disease; and environmental disease. The values for the incidence of these disorders are only approximate estimates. Overall, complex disorders are far more common than single-gene disorders. However, it is far easier to discover the genetic defect that underlies single-gene disorders. For all categories of disease, the pathophysiology (i.e. the disease-altered physiological processes) depends on the influence of many genes.*

CATEGORIES OF DISEASE

What kinds of diseases afflict humans? How can diseases be classified? We can begin with a list of five main categories (Figure 18-2). In a basic way, these categories are arbitrary because every disease is manifested with many phenotypes in different individuals, and every disease involves many genes.

1. Mendelian disorders, also called single-gene or monogenic disorders, are caused primarily by mutations in single genes. The Mendelian disorders that Garrod described in his 1909 book occur only rarely in the general population. Overall Mendelian disorders occur in about 1% of the general population, and are almost always manifested early in life (before adulthood). In these diseases, mutations in one gene have a major influence on the disease. However, even in single-gene disorders, many additional genes may function as modifiers, affecting the disease phenotype. Thus two individuals with the identical mutation in a gene may have entirely different patterns of expression of a disease (discussed below).
2. Complex disorders such as Alzheimer's disease and cardiovascular disease are caused by defects in multiple genes. These disorders are also called multifactorial, reflecting that they are expressed as a function of both genetic and environmental factors. In comparison to monogenic disorders, complex

See Table 18.22 (p. 689) for a list of cancer resources.

disorders tend to be highly prevalent. In the United States, chronic diseases such as heart disease, senile dementia, cancer, and diabetes are the leading causes of death and disability. These all have some degree of genetic basis.

3. Chromosomal disorders are arguably the most common causes of disease, from the perspective that lethal chromosomal abnormalities occur in perhaps 50% to 75% of all human conceptuses (see page 673 below).
4. From birth to old age, infectious disease is the leading cause of death worldwide.
5. Environmental disease, defined as having a non-genetic proximal cause, are also extremely common. For example, infectious disease and environmental toxins can cause disease independent of a mechanism involving a genetic mutation.

There are various causes of human disease, whether genetic or not. Consistent with Garrod's perspective, the pathophysiology of any disease may be considered multigenic. Two individuals who are exposed to the same disease-causing stimulus—whether it is a virus or lead paint or a mutated gene—may have entirely different reactions. One person may become ill, while the other is unaffected. There is a large genetic component to the responses to any disease-causing condition.

CLASSIFICATION OF DISEASE

We described several general categories of disease in Figure 18.2. From the perspective of bioinformatics, we are interested in understanding the mechanism of disease in relation to genomic DNA, genes, and their gene products. We are further interested in the consequences of mutations on cell function and on the comparative genomics of disease-causing genes throughout evolution. This perspective is quite different than that of the clinician or epidemiologist.

There are many approaches to classifying disease. One is to describe mortality statistics. These data (based on death certificates in the United States) include rankings of the cause of death (Table 18.2) (Anderson, 2001). This information is helpful in identifying the most common diseases.

Lead poisoning is an example of a disease caused by an environmental agent. About 8% of all children in the United States have blood levels that are defined as "alarming," according to the Centers for Disease Control and Prevention. See ▸http://www.cdc.gov/mmwr/preview/mmwrhtml/mm4950a3.htm#tabl.

Pathology is the study of the nature and cause of disease. Pathophysiology is the study of how disease alters normal physiological processes.

Ranking the cause of death is a somewhat arbitrary procedure that depends on the list of causes from which the selection is made and the use of a uniform ranking procedure (Anderson, 2001). For example, the list in Table 18.2 does not include lung cancer (a subset of neoplasms) or motor vehicle accidents (a subset of accidents), each of which would otherwise rank in the top 10.

The data in Table 18.2 are available from the National Center for Health Statistics. Their website is at ▸http://www.cdc.gov/nchs/.

TABLE 18-2 Leading Causes of Death in United States (1999)

Rank	Cause	Number of Deaths	Percent of Total Deaths
1	Diseases of heart	725,192	30.3
2	Malignant neoplasms	549,838	23.0
3	Cerebrovascular disease	167,366	7.0
4	Chronic lower respiratory diseases	124,181	5.2
5	Accidents	97,860	4.1
6	Diabetes mellitus	68,399	2.9
7	Influenza and pneumonia	63,730	2.7
8	Alzheimer's disease	44,536	1.9
9	Nephritis, nephritic syndrome, nephrosis	35,525	1.5
10	Septicemia	30,680	1.3
	All other causes	484,092	20.2
	All causes	2,391,399	100.0

Source: Modified from R. N. Anderson, *National Vital Statistics Reports* **49**(11); 1-87 (2001). Used with permission.

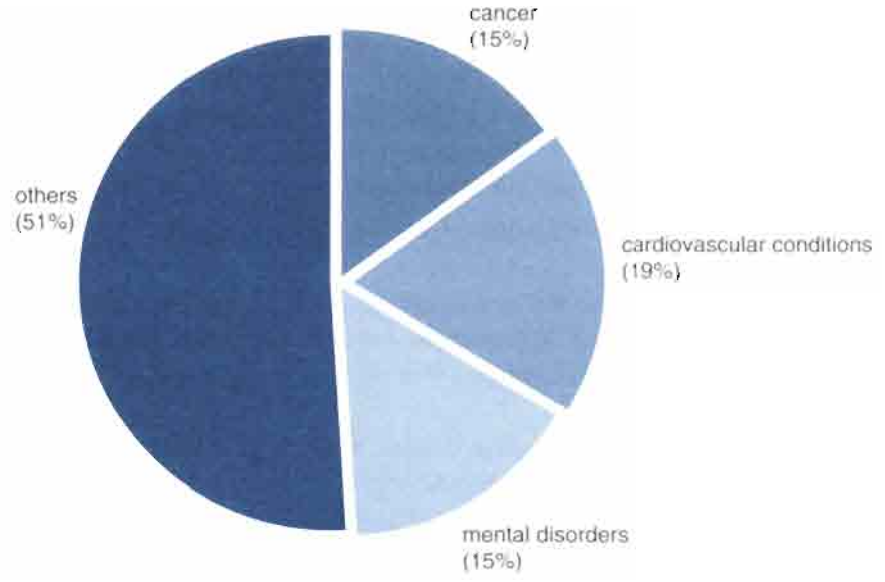


FIGURE 18.3. The global burden of disease. Adapted from ▸ *http://www.surgeongeneral.gov/library/mentalhealth/chapter6/sec2.html#figure6_3.*

A summary of the Global Burden of Disease report is available at ▸ http://www.who.int/msa/mnh/ems/dalys/intro.htm. DALYs are calculated by adding the years of life lost through all deaths in a year plus the years of life expected to be lived with a disability for all cases beginning in that year.

Another approach to describing the scope of human disease is to measure the global burden of disease in terms of the percentage of affected individuals (Fig. 18.3) or in terms of disability-adjusted life years (DALYs) (Fig. 18.4). Worldwide, noncommunicable diseases such as depression and heart disease are rapidly replacing infectious diseases and malnutrition as the leading causes of disability and premature death (Murray and Lopez, 1996).

A far more extensive listing of morbidity data is provided by the International Statistical Classification of Diseases and Related Health Problems (abbreviated

Rank	cause (1990)	% total
1	Lower respiratory infections	8.2%
2	Diarrheal diseases	7.2%
3	Perinatal conditions	6.7%
4	Unipolar major depression	3.7%
5	Ischemic heart disease	3.4%
6	Cerebrovascular disease	2.8%
7	Tuberculosis	2.8%
8	Measles	2.7%
9	Road traffic accidents	2.5%
10	Congenital abnormalities	2.4%

Rank	cause (2020)	% total
1	Ischemic heart disease	5.9%
2	Unipolar major depression	5.7%
3	Road traffic accidents	5.1%
4	Cerebrovascular disease	4.4%
5	Chronic obs pulmonary dis.	4.4%
6	Lower respiratory infections	3.1%
7	Tuberculosis	3.0%
8	War	3.0%
9	Diarrheal diseases	2.7%
10	HIV	2.6%

FIGURE 18.4. Global disease burden measured in DALYS for 1990 and for 2020 (projected). In females and in developing countries, unipolar major depression is projected as the leading cause of disease burden. From Murray and Lopez (1996), p. 375. Used with permission.

TABLE 18-3 ICD Classification System

1. Infectious and parasitic diseases
2. Neoplasms
3. Endocrine, nutritional, and metabolic diseases and immunity disorders
4. Diseases of the blood and blood-forming organs
5. Mental disorders
6. Diseases of the nervous system and sense organs
7. Diseases of the circulatory system
8. Diseases of the respiratory system
9. Diseases of the digestive system
10. Diseases of the genitourinary system
11. Complications of pregnancy, childbirth, and the puerperium
12. Diseases of the skin and subcutaneous tissue
13. Diseases of the musculoskeletal system and connective tissue
14. Congenital anomalies
15. Certain conditions originating in the perinatal period
16. Symptoms, signs, and ill-defined conditions
17. Injury and poisoning

Source: From ICD-9 as described in the KEGG database, ▸http://www.genome.ad.jp/dbget-bin/get_htext?ICD9.

ICD). This resource, published by the World Health Organization (WHO), is used to classify diseases (Table 18.3).

Mortality statistics list the most common diseases. We are interested in the full spectrum of disease, including rare diseases. These are defined as diseases affecting fewer than 200,000 people. In the United States, an estimated 25 million individuals (almost 10% of the population) suffer from one or more of 6000 rare diseases.

The WHO ICD website is at ▸http://www.who.int/whosis/icd10/ This resource was begun in 1893 as the International List of Causes of Death.

The Office of Rare Diseases at the National Institutes of Health (NIH) has a website that lists hundreds of rare diseases (▸http://ord.aspensys.com/diseases.asp).

NIH Disease Classification: MeSH Terms

The National Library of Medicine (NLM) has developed Medical Subjects Heading (MeSH) terms as a unified language for biomedical literature database searches. The 2003 MeSH term system includes 23 disease categories (Fig. 18.5). PubMed at NCBI also uses this classification system.

MeSH terms are a controlled vocabulary used to index MEDLINE (and PubMed, which is based on MEDLINE). A search for the term "Rett Syndrome" at the NLM MeSH site shows the hierarchical tree structure of the MeSH terms: Rett syndrome is listed separately under categories such as mental retardation, neurodegenerative disorders, and inborn genetic diseases.

We will discuss Rett syndrome below. You can access the MeSH system at NLM (▸http://www.nlm.nih.gov/mesh/meshhome.html) or at NCBI (from PubMed, select MeSH terms on the left sidebar, then enter the query "disease."

MONOGENIC DISORDERS

Our perspectives on the molecular nature of disease have evolved in recent decades. Previously, geneticists recognized a dichotomy between simple traits and complex traits. More recently, all traits have come to be appreciated as part of a continuum.

A trait is a characteristic or property of an individual that is the outcome of the action of a gene or genes.

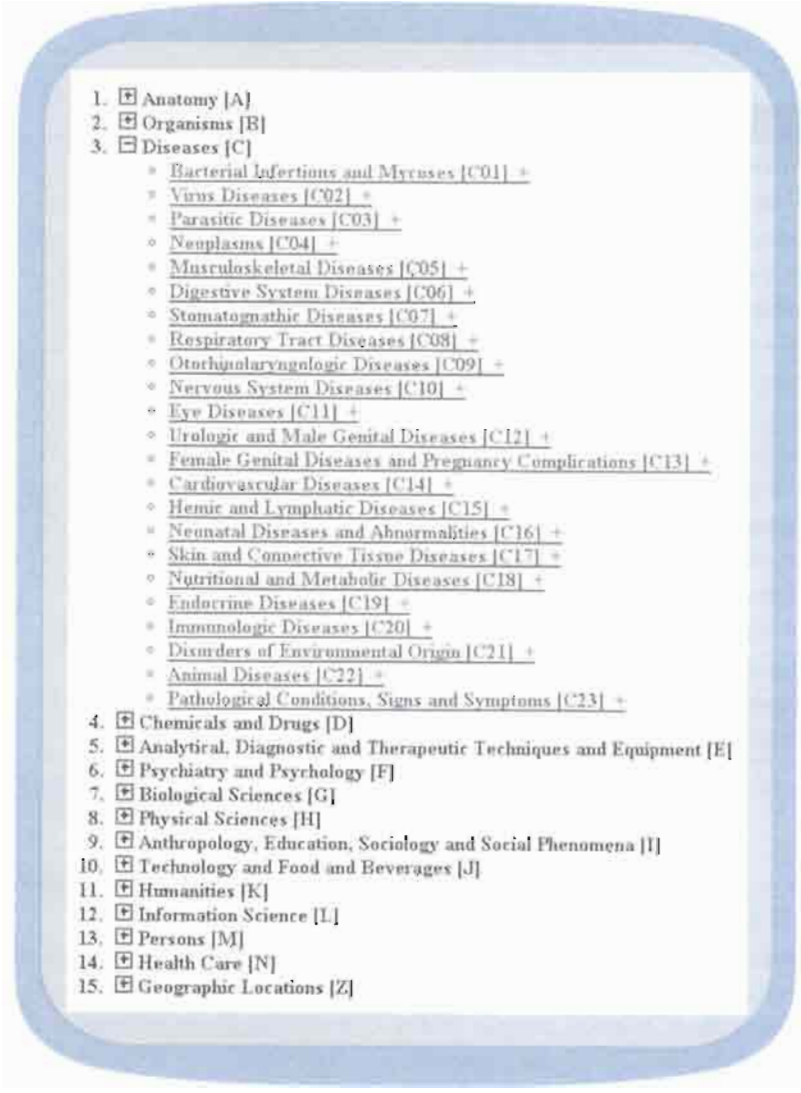


FIGURE 18.5. *The Medical Subject Heading (MeSH) term system at the National Library of Medicine includes 15 major categories (2003 version). The disease category further includes the 23 headings shown here. See* ▸*http://www.nlm.nih.gov/mesh/.*

TABLE 18-4 Examples of Monogenic Disorders

Mechanism	Disorder	Frequency
Autosomal dominant	*BRCA1* and *BRCA2* breast cancer	1 in 1000 (1 in 100 for Ashkenazim)
	Huntington chorea	1 in 2500
	Neurofibromatosis I	1 in 3000
	Tuberous sclerosis	1 in 15,000
Autosomal recessive	Albinism	1 in 10,000
	Sickle cell anemia	1 in 655 (U.S. African-Americans)
	Cystic fibrosis	1 in 2500 (Europeans)
	Phenylketonuria	1 in 12,000
X linked	Hemophilia A	1 in 10,000 (males)
	Glucose 6-phosphate dehydrogenase deficiency	Variable; up to 1 in 10 males
	Fragile X syndrome	1 in 1250 males
	Color blindness	1 in 12 males
	Rett syndrome	1 in 20,000 females
	Adrenoleukodystrophy	1 in 17,000

Source: Adapted from Beaudet et al. (2001). Used with permission.

BOX 18-1
Sickle Cell Anemia

Our cells depend on oxygen to live, and blood transports oxygen throughout the body. However, oxygen is a hydrophobic molecule that requires the carrier protein hemoglobin to transport it through blood. (The homologous protein myoglobin transports oxygen in muscle cells.) Adult hemoglobin is composed of two α chains and two β chains. Mutation in the β chain (NM_000518 and NP_000509) on chromosome 11p15.5 causes sickle cell anemia. Red blood cells in patients can assume a curved, "sickled" appearance that reflects hemoglobin aggregation in the presence of low oxygen levels.

Sickle cell anemia is the most common inherited blood disorder in the United States, affecting 1 in 500 African-Americans. It is inherited as an autosomal recessive disease. Heterozygotes (individuals with one normal copy of hemoglobin beta and one mutant copy) are somewhat protected against the malaria parasite, *Plasmodium falciparum*. This may be because normal red blood cells infected by the parasite are destroyed. Thus there is a selective evolutionary pressure to preserve the HBS mutation in the population that is at risk for malaria.

Following are web resources for sickle cell disease:

Resource	URL
NIH fact sheet	▸http://www.nhlbi.nih.gov/health/public/blood/sickle/sca_fact.pdf
Genes and Disease (NCBI)	▸http://www.ncbi.nlm.nih.gov/disease/sickle.html
Sickle Cell Disease Association of America	▸http://www.sicklecelldisease.org/

We examined the structure of normal beta globin (HBB) as well as the most common mutated form (HBS) in Chapter 9. The E6V substitution (valine in place of glutamate as the sixth amino acid) adds a hydrophobic patch to the protein, promoting the aggregation of globin molecules and the formation of sickle-shaped red blood cells.

You can read the Pauling et al. (1949) ariticle on-line at ▸http://profiles.nlm.nih.gov/MM/B/B/R/L/. The National Library of Medicine (NLM) offers on-line access to all the publications of several prominent biologists through its Profiles in Science site (▸http://profiles.nlm.nih.gov/). The scientists include Linus Pauling and other Nobel Prize laureates such as Barbara McClintock, Julius Axelrod, and Oswald Avery.

Sickle cell anemia is unusually common for a single-gene disorder. This is presumably because of the protection it confers to heterozygotes exposed to malaria (Box 18.1).

Simple traits are transmitted following the rules of Mendel. Several monogenic disorders are listed in Table 18.4. As an example of a single-gene disorder, consider sickle cell anemia (Box 18.1). In 1949 Linus Pauling and colleagues described the abnormal electrophoretic behavior of sickle cell hemoglobin (Pauling et al., 1949). It was subsequently shown that a single amino acid substitution accounts for the abnormal behavior of the sickle cell and is the basis of sickle cell anemia. This is a single-gene disorder that is inherited in an autosomal recessive fashion. Single-gene disorders tend to be rare in the general population. Note that sickle cell disease is the outcome of having a particular mutant hemoglobin protein. While there are common features of sickle cell disease, such as sickling of the red blood cells, there is not a single disease phenotype. The pleiotropic phenotype is caused by the influence of other genes.

Rett syndrome is another example of a single-gene disorder (Box 18.2). This disease affects girls almost exclusively. While they are apparently born healthy, Rett syndrome girls acquire a constellation of symptoms beginning at 6–18 months of age. They lose the ability to make purposeful hand movements, and they typically exhibit hand-wringing behavior. Whatever language skills they have acquired are lost, and they may display autistic-like behaviors. Rett syndrome is caused by mutations in the gene encoding MeCP2, a transcriptional repressor that binds methylated CpG islands (Amir et al., 1999) (see Chapter 17). It is not yet known why mutations affecting a transcriptional repressor that functions throughout the body cause a primarily neurological disorder.

BOX 18-2
Rett Syndrome

Rett syndrome (RTT; OMIM #312750) is a developmental neurological syndrome that occurs almost exclusively in females (Hagberg et al., 1983; Armstrong, 1997; Naidu, 1997). Affected females are apparently normal through pre- and perinatal development, following which there is a developmental arrest. This is accompanied by decelerated head and brain growth, loss of speech and social skills, severe mental retardation, truncal ataxia, and characteristic hand-wringing motions. Prominent neuropathological features include reductions in cortical thickness in multiple cerebral cortical regions, reduced neuronal soma size, and dramatically decreased dendritic arborization (Jellinger et al., 1988; Bauman et al., 1995; Belichenko et al., 1997).

Mutations in the methyl-CpG-binding protein 2 (*MECP2*) gene located in Xq28 have been found in many cases of RS (Amir et al., 1999; Wan et al., 1999; Amir et al., 2000; Bienvenu et al., 2000; Cheadle et al., 2000; Huppke et al., 2000; Kim and Cook, 2000; Van den Veyver and Zoghbi, 2000; Xiang et al., 2000). MeCP2 binds to methylated CpG dinucleotides throughout the genome and is involved in methylation-dependent repression of gene expression via the recruitment of the corepressor mSin3A, and the chromatin remodeling histone deacetylases HDAC1 and HDAC2. The expression of MeCP2 mRNA in many tissues and its interaction with regulatory DNA elements in multiple chromosomes suggest that MeCP2 is a global repressor of gene expression (Nan et al., 1997). DNA methylation-dependent repression of gene expression has been associated with genetic imprinting, X-chromosome inactivation, carcinogenesis, and tissue-specific gene expression (Cedar, 1988; Allaman-Pillet et al., 1998; Razin, 1998; Ng and Bird, 1999).

Why is it that some tissues are spared the effects of *MECP2* mutations? There could be tissue-specific redundancy of gene function, or other compensation mechanisms. This provides an example of how the tools of bioinformatics are relevant to studying many different aspects of human disease.

While Rett syndrome is a disease caused by a mutation in a single gene, it exemplifies the extraordinary complexity of human disease and even monogenic disorders:

- The disease occurs primarily in females. It was thought that this could be explained by the location of the *MECP2* gene on the X chromosome: A mutation in this gene might be lethal for males in utero (having only a single X chromosome), while females might have the disease phenotype because they have one normal and one mutant copy of the gene. Instead, the more likely explanation is that most mutations occur in fathers. The father is healthy, but a new germline mutation arises and is passed to daughters. Thus all sons (XY) receive a normal Y chromosome from the father, while daughters receive a mutant copy of the X chromosome from the father.
- After the discovery that mutations in *MECP2* cause Rett syndrome, it was discovered that some males with mental retardation also have mutations in this gene (Hammer et al., 2002; Geerdink et al., 2002; Zeev et al., 2002). However, the phenotype of mutations in the male is distinctly different than in females, often involving severe neonatal encephalopathy. In males, having a single X chromosome, the mutant gene is expected to adversely affect virtually every cell in the body. In contrast, females undergo random X-chromosome inactivation: Having two copies of the X chromosome, every cell expresses only one chromosome (either the maternal or paternal chromosome, randomly selected early in development). Thus females are a mosaic in terms of X-chromosome allelic expression, and a Rett syndrome female has on average 50% normal cells throughout her body.

- While Rett syndrome is caused by mutations in a gene encoding a transcriptional repressor, it is almost certain that the consequence of this mutation involves subsequent effects on the expression of many other genes. Thus, like any other monogenic disorder, many other genes are involved and may influence the phenotype of the disease.
- Two females having the identical mutation in *MECP2* may have entirely different phenotypes (in terms of severity of the disease). There are two main explanations for this observation, which is seen for many other single-gene disorders as well. (1) There may be modifier genes that influence the disease process (Dipple and McCabe, 2000). Modifier genes have been identified for patients with sickle cell anemia, adrenoleukodystrophy, cystic fibrosis, and Hirschprung disease. Most (if not all) apparently monogenic disorders are complex. (2) A variety of epigenetic influences may drastically affect the clinical phenotype. For example, the methylation status of genomic DNA could determine the molecular consequences of mutations in *MECP2*. X chromosome inactivation is sometimes skewed, such that the phenotype is more severe (if the X chromosome copy with mutant *MECP2* is preferentially expressed) or less severe (if the healthy X chromosome is selectively expressed).

OMIM: Central Bioinformatics Resource for Human Disease

Mendelian Inheritance in Man (MIM) was started in 1966 by Victor A. McKusick. The online version OMIM became integrated with NCBI in 1995. It is available at ▸http://www.ncbi.nlm.nih.gov/omim/ or preferably through Entrez at NCBI.

In considering Mendelian disorders, we turn now to OMIM, a comprehensive database for human genes and genetic disorders (Hamosh et al., 2000, 2002). The OMIM database contains bibliographic entries for over 12,000 human diseases and relevant genes. The focus of OMIM is inherited genetic diseases. As indicated by its name, the OMIM database is concerned with Mendelian genetics. These are inherited traits that are transmitted between generations. There is relatively little known about genetic mutations in complex disorders (see below), and the database does not include chromosomal disorders (see below). Thus its focus is a comprehensive survey of single-gene disorders.

We can examine OMIM using Rett syndrome and *MECP2* as an example of a disease and a gene implicated in a disease. OMIM can be searched from the NCBI Entrez site, and it is linked from LocusLink. Within the OMIM site, there is a search page that allows you to query a variety of fields, including chromosome, map position, or clinical information (Fig. 18.6). The result includes both the disease Rett syndrome and the relevant gene, *MECP2* (Fig. 18.7). Since mutations in *MECP2* are associated with distinct phenotypes, these are also listed (e.g., Mental retardation, X-linked). Another related disease is listed, Angelman syndrome, because some patients with mutations in *MECP2* have some of the clinical features of that disease.

We can next view the entry for Rett syndrome (Fig. 18.8), with its OMIM identifier #312750. Each entry in OMIM is associated with a numbering system. There is a six-digit code in which the first digit indicates the mode of inheritance of the gene involved (Table 18.5). The Rett syndrome entry is preceded by a number sign to indicate that the phenotype can be caused by mutations in more than one gene. The first number (3) indicates that this disorder is X linked.

The OMIM entry for *MECP2* (*300005) includes an asterisk, indicating that the phenotype determined by the gene at the given locus is separate from those

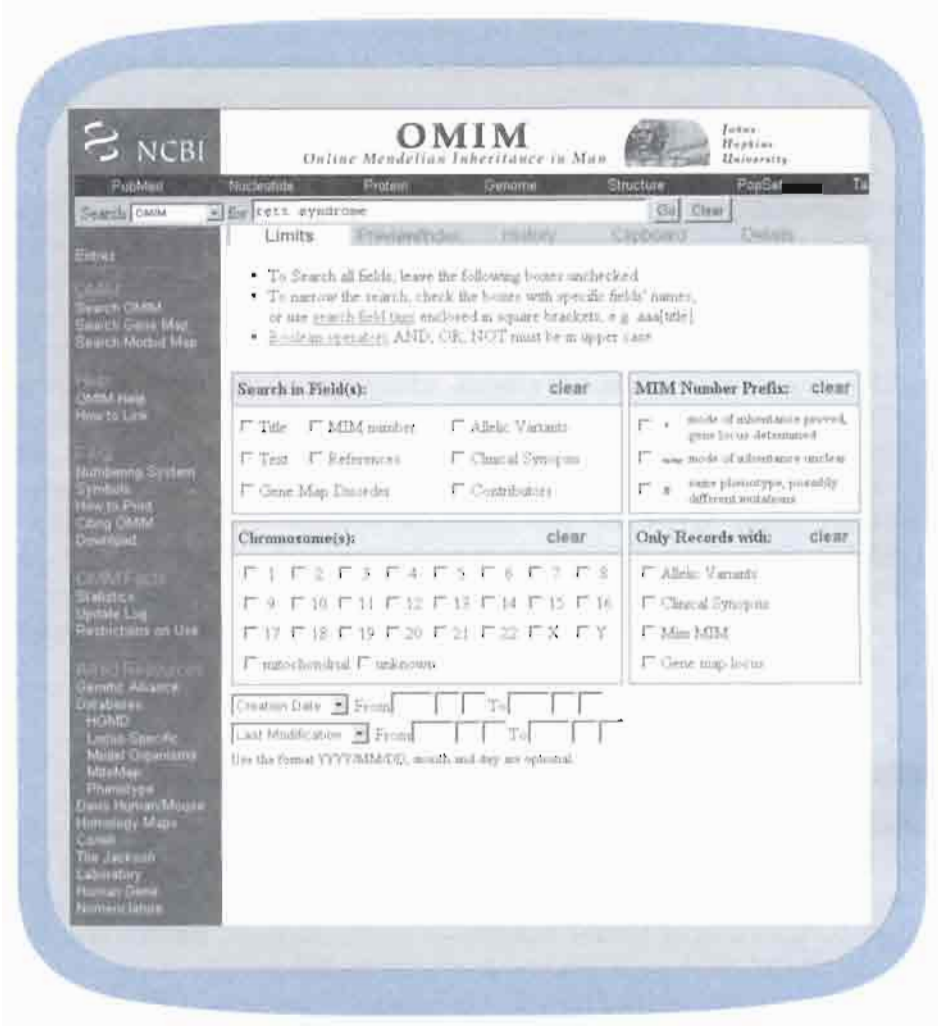


FIGURE 18.6. *Online Mendelian Inheritance in Man (OMIM), accessible via the NCBI website (http://www.ncbi.nlm.nih.gov/omim), allows text searches by criteria such as author, gene identifier, or chromosome.*

Allelic variants are selected based on criteria such as being the first mutation to be discovered, having a high population frequency, or having an unusual pathogenetic mechanism. Some allelic variants in OMIM represent polymorphisms. These may be of particular interest if they show a positive correlation with common disorders (see below).

To find a list of human genes which are associated with disease, go to LocusLink, restrict the search to human, and enter the query "disease_known AND has_seq." There are currently 1521 entries for human inherited diseases having a known sequence (June 2003).

phenotypes represented by other asterisked entries and that the mode of inheritance of the phenotype has been proved. The entry includes bibliographic data such as available information on an animal model for Rett syndrome.

OMIM entries link to a gene map, which provides a tabular listing of the cytogenetic position of disease loci (Fig. 18.9). This gene map further links to the NCBI Map Viewer (Fig. 18.9, arrow 1) and to resources for the orthologous mouse gene. The OMIM morbid map also provides cytogenetic loci but is organized alphabetically. The current holdings of OMIM, arranged by chromosome, are given in Table 18.6.

An important feature of OMIM entries is that many contain a list of allelic variants. Most of these represent disease-causing mutations. An example of several allelic variant entries is shown for *MECP2* (Fig. 18.10). These allelic variants provide a glimpse of all the human genes that are known to contain disease-causing mutations.

The current holdings of OMIM are summarized in Table 18.7. OMIM continues to be a crucial and comprehensive resource for information on the human genome. The database is maintained and updated by expert curators (Hamosh et al., 2002). Many other disease databases incorporate OMIM identifiers to provide a common reference to disease-related genes.

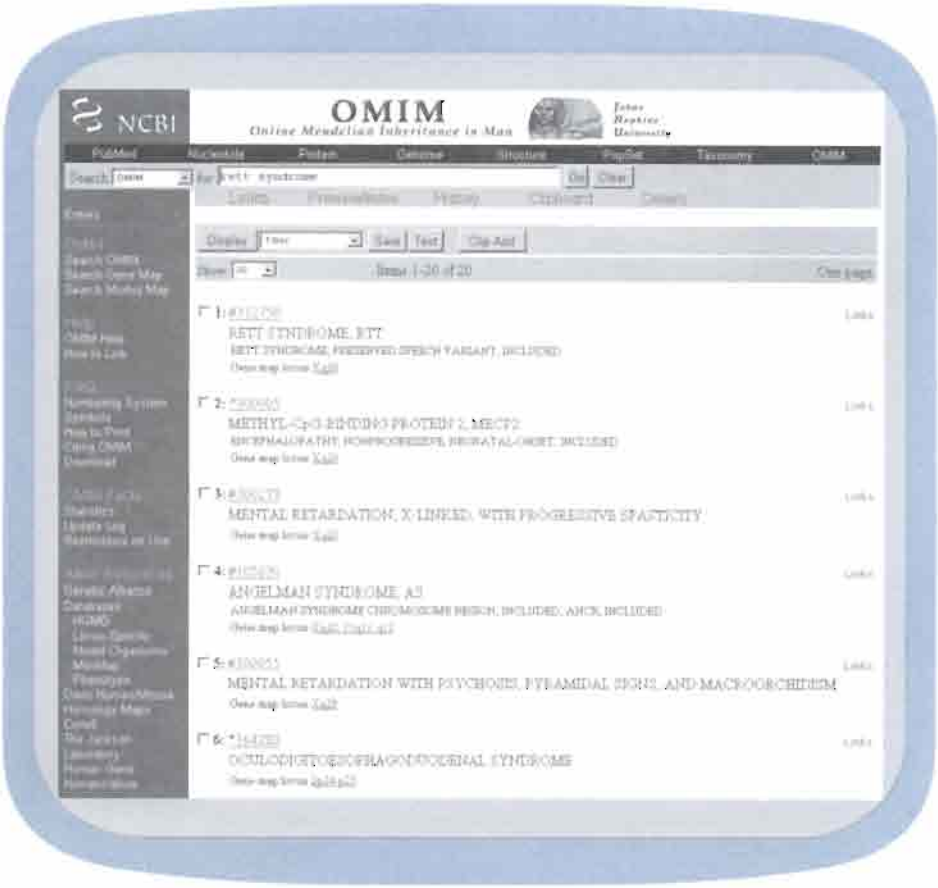


FIGURE 18.7. *A search of OMIM for "Rett syndrome" produces entries on that syndrome as well as related entries on* MECP2 *(the gene that, when mutated, causes Rett syndrome), entries on mental retardation syndromes with distinct phenotypes that are caused by mutations in the same gene (e.g., X-linked mental retardation with progressive spasticity), and distinct disorders such as Angelman syndrome that are associated with a closely similar phenotype.*

Other Central Human Disease Databases

There are other databases that provide a comprehensive view of human disease genes. An example is GeneCards (Safran et al., 2002). This is a human gene compendium that includes a wealth of information on human disease genes. GeneCards differs from OMIM in that it collects and integrates data from several dozen independent databases including OMIM, GenBank, UniGene, Ensembl, the University of California at Santa Cruz (UCSC), and the Munich Information Center for Protein Sequences (MIPS). Thus, relative to OMIM, GeneCards uses relatively less descriptive text of human diseases, and it provides relatively more functional genomics data. GeneCards also uses Human Genome Organisation (HUGO) gene symbols to ensure a standard nomenclature. For *MECP2*, the GeneCards entry includes data on the chromosomal locus, SwissProt links, regional gene expression data based on microarray experiments and UniGene statistics, and orthologs (Fig. 18.11). Additional central resources are listed in Table 18.8. Another central database is the Human Gene Mutation Database (HGMD) (Krawczak and Cooper, 1997). This database collates published gene alterations responsible for human inherited disease.

Mutation Databases

There are two basic kinds of human disease gene databases: central and locus-specific. Central databases such as OMIM, GeneCards, and HMG attempt to

As an example of OMIM links, the TIGR Eukaryotic Gene Ortholog (EGO) database (▸http://www.tigr.org/) includes a link of each of its Human Gene Index entries to OMIM at ▸http://www.tigr.org/tdb/tgi/ego/human_dis_gene.shtml. Similarly, the Ensembl (▸http://www.ensembl.org) and UCSC (▸http://genome.ucsc.edu) genome browsers (Chapter 17) link to OMIM.

GeneCards is available at ▸http://bioinfo.weizmann.ac.il/cards/. It is "a database of human genes, their products and their involvement in diseases."

HGMD is available at ▸http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html. It contains 30,641 mutations in 1245 genes and provides 1100 reference cDNA sequences (September 2002 release).

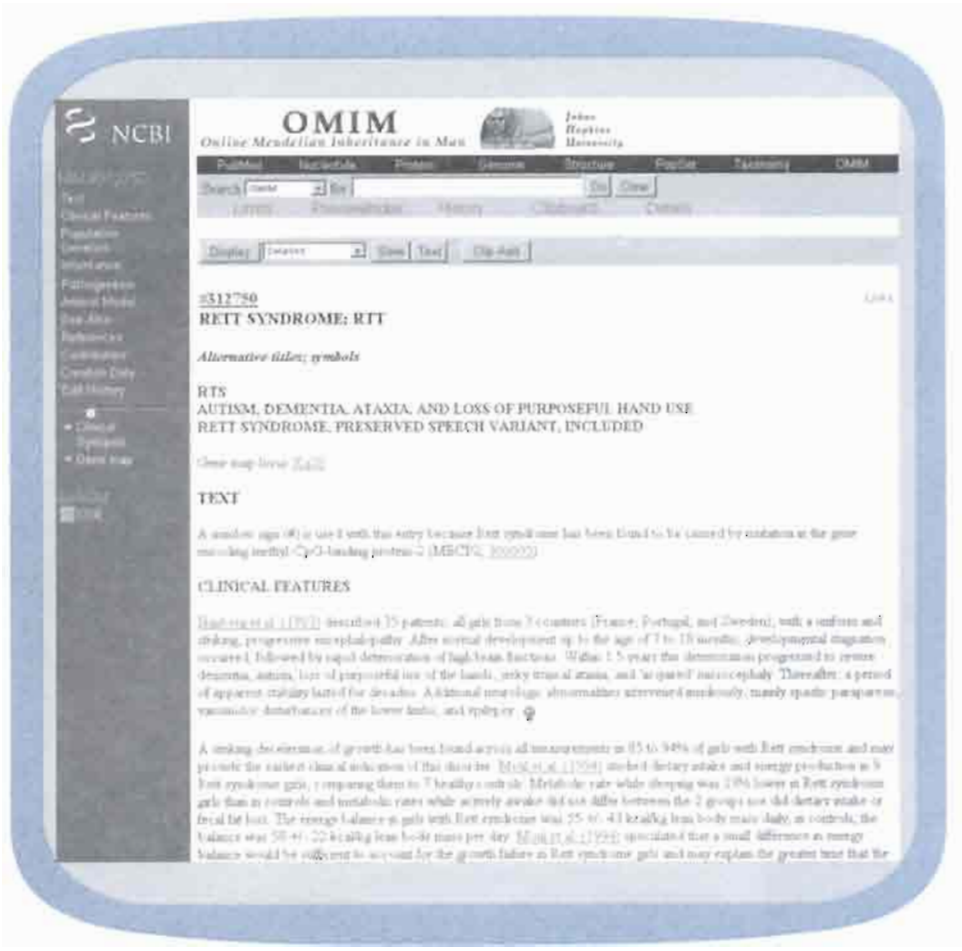


FIGURE 18.8. *The OMIM entry for Rett syndrome includes the OMIM identifier (#312750) and a variety of information, listed on the sidebar, such as clinical features and a description of available animal models.*

TABLE 18-5 OMIM Numbering System

The entries beginning 1 and 2 entered the database before May 1994; those beginning with 6 were created after May 1994

OMIM Number	Phenotype	OMIM Identifier	Disorder (example)	Chromosome
1___	Autosomal dominant	*143100	Huntington disease	4p16.3
2___	Autosomal recessive	*209850	Autism, susceptibility to, (AUTS1)	7q
3___	X-linked loci or phenotypes	#312750	Rett syndrome	Xq28
4___	Y-linked loci or phenotypes	*480000	Sex-determining region Y	Yp11.3
5___	Mitochondrial loci or phenotypes	#556500	Parkinson disease	–
6___	Autosomal loci or phenotypes	#603903	Sickle cell anemia	–

Note. An asterisk (*) preceding an entry indicates that the phenotype determined by the gene at the given locus is separate from those represented by other asterisked entries, and the mode of inheritance of the phenotype has been proved (in the judgment of the authors and editors). Generally, there is only one asterisked entry per gene locus. No asterisk before an entry number means that the mode of inheritance has not been proved, although suspected, or that the separateness of this locus from that of another entry is unclear. A number symbol (#) indicates that the phenotype can be caused by mutation in any of two or more genes. For the AUTS1 entry, the number 1 indicates that this is the first listing of several autism susceptibility loci (e.g. AUTS2).

Source: Adapted from ▸http://www.ncbi.nlm.nih.gov:80/entrez/Omim/ominfaq.html#numbering.system.

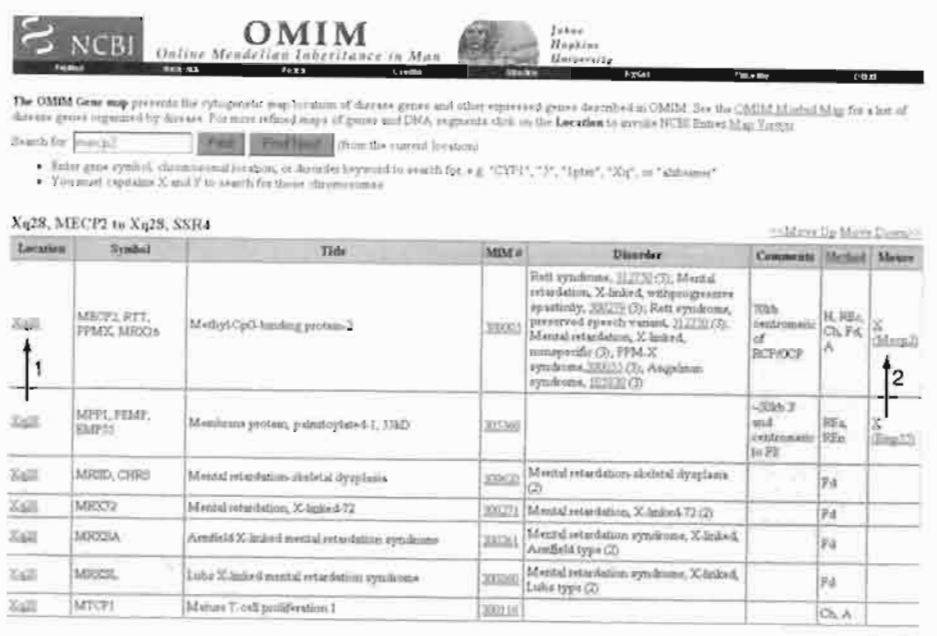


FIGURE 18.9. *The OMIM entry for Rett syndrome links to the OMIM gene map. This map provides a tabular listing of the cytogenetic position of disease loci. Clicking on the location link (arrow 1) leads to the Entrez map viewer (Fig. 17.2). The mouse link (arrow 2) leads to the* MECP2 *page at the Mouse Genome Informatics site (▸http://www.informatics.jax.org).*

comprehensively describe all disease-related genes. In contrast, locus-specific databases—also called mutation databases—describe variations in a single gene (or sometimes in several genes) in depth. These databases provide particular expertise on the genetic aspects of one specific gene. Also, the coverage of known mutations tends to be far deeper in locus-specific databases as a group than in central databases (Scriver et al., 1999). Thus, these two types of databases serve complementary purposes.

A locus-specific database is a repository for allelic variations. There are hundreds of such databases. The essential components of a locus-specific database include the following (Scriver et al., 1999, 2000; Claustres et al., 2002):

- A unique identifier for each allele
- Information on the source of the data
- The context of the allele
- Information on the allele (e.g., its name, type, and nucleotide variation)

In the context of mutation databases, a mutation is defined as an allelic variant (Scriver et al., 1999). The allele (or the unique sequence change) may also be disease causing; such an allele tends to occur at low frequency. The allele may also be neutral, not having any apparent effect on phenotype.

TABLE 18-6 Synopsis of OMIM Human Genes per Chromosome (November 2002)

See ▸http://www.ncbi.nlm.nih.gov/entrez/Omim/mimstats.html.

Chromosome	Loci	Chromosome	Loci	Chromosome	Loci
1	795	9	294	17	474
2	515	10	271	18	120
3	432	11	516	19	537
4	305	12	421	20	185
5	413	13	138	21	114
6	484	14	252	22	189
7	382	15	234	X	493
8	287	16	308	Y	34

Note: Total number of loci: 8193

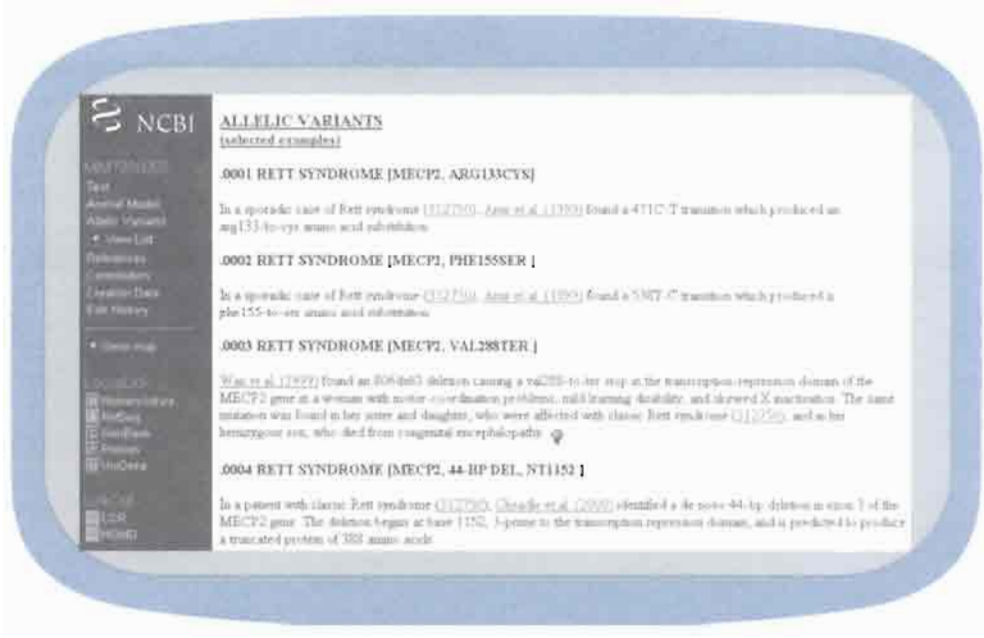


FIGURE 18.10. *The OMIM entry for* MECP2 *includes allelic variants, most of which are disease-causing mutations. Some allelic variants reflect polymorphisms that are not associated with disease.*

Mutation databases have an important role in gathering information about mutations, but there have not been uniform standards for their creation until recently. Claustres et al. (2002) surveyed 94 websites that encompassed 262 locus-specific databases. They found great variability in the way data are collected, presented, linked, named, and updated. Scriver et al. (1999, 2000) described guidelines for the content, structure, and deployment of mutation databases.

- There is now increased uniformity in naming alleles (Antonarakis, 1998; den Dunnen and Antonarakis, 2000). For example, the A of the ATG of the initiator Met codon is denoted nucleotide +1. Many such rules have been explicitly stated to allow uniform descriptions of mutations.
- Ethical guidelines have been described, such as the obligation of preserving the confidentiality of information (Knoppers and Laberge, 2000).
- Generic software to build and analyze locus-specific databases has been provided (Beroud et al., 2000; Brown and McKie, 2000).

To see the Universal Mutation Database template of Beroud et al. (2000), visit ▸http://www.umd.necker.fr/.

Several websites provide gateways to access locus-specific databases (Table 18.9). The main point of entry is the Human Genome Variation Society (HGVS) (Fig. 18.12). This provides access to hundreds of locus-specific mutation databases.

HGVS is accessible at ▸http://ariel.ucs.unimelb.edu.au:80/~cotton/mdi.htm. It was formerly called the HUGO Mutation Database Initiative. HGVS is mirrored at the EMBL-EBI (▸http://www2.ebi.ac.uk/mutations/cotton/).

TABLE 18-7 Current Holdings of OMIM (November 2002)

See ▸http://www.ncbi.nlm.nih.gov/entrez/Omim/mimstats.html.

	Autosomal	X Linked	Y Linked	Mitochondrial	Total
Established genes or phenotype loci (*)	9,792	538	39	37	10,406
Phenotype descriptions (#)	1,065	90	0	23	1,178
Other loci or phenotypes (no prefix)	2,263	158	2	0	2,422
Total	13,120	786	41	60	14,017

Note: See Table 18.5 note for definition of the asterisk (*) and number (#) symbols.

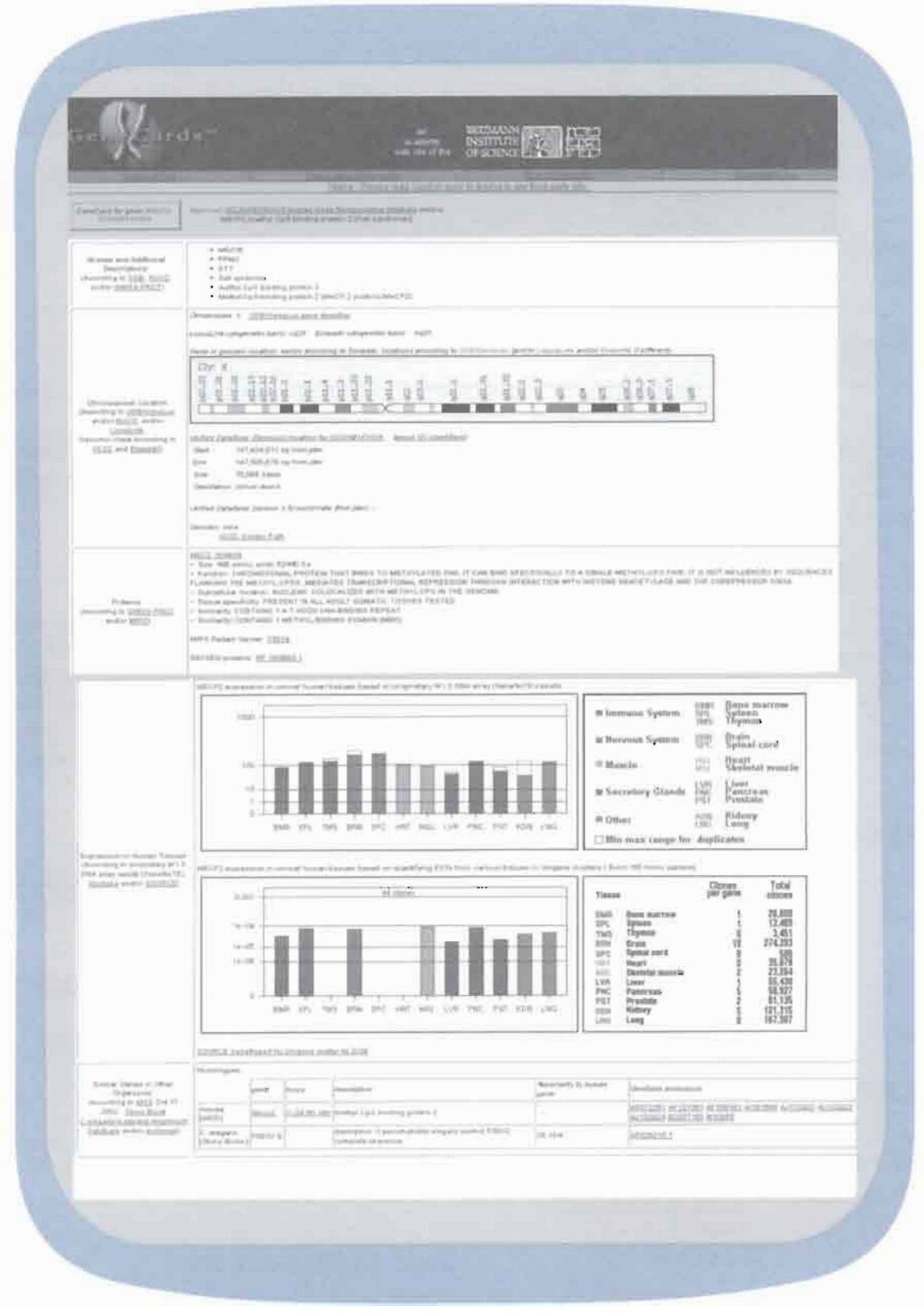


FIGURE 18.11. *GeneCards is a genomics database that includes a wealth of data on gene sequences, genomic locations, gene expression, protein sequence and structure, and medical information. Its URL is* ▸*http://bioinfo.weizmann.ac.il/cards/.*

TABLE 18-8 Central Mutation Databases on World Wide Web

Site	Description	URL
GeneCards	At the Weizmann Institute	▸http://bioinfo.weizmann.ac.il/cards/
HUGO Mutation Database Initiative	A large list of locus-specific mutation databases	▸http://ariel.ucs.unimelb.edu.au:80/~cotton/glsdb.htm
Human Gene Mutation Database	From the Institute of Medical Genetics in Cardiff	▸http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html
The Mammalian Gene Mutation Database (MGMD)	Database of published mutagen-induced gene mutations in mammalian tissues	▸http://lisntweb.swan.ac.uk/cmgt/index.htm
Sequence Variation Database (SVD) project	At EBI	▸http://www2.ebi.ac.uk/mutations/

RettBASE (▸http://mecp2.chw.edu.au/) is curated by John Christodoulou and Andrew Grimm (Children's Hospital, Westmead, Sydney). It was supported by the International Rett Syndrome Association (IRSA). Founded by parents of Rett syndrome girls, IRSA and the Rett Syndrome Research Foundation are two groups that promote research and education into Rett syndrome. Similar advocacy groups exist for almost all human diseases (see below).

As an example of a locus-specific database, we can examine RettBASE, an *MECP2* database. RettBASE provides detailed data on mutations in *MECP2* and the corresponding clinical phenotype (Fig. 18.13).

Single-Nucleotide Polymorphisms and Disease

In a comparison of any two randomly selected human genomes of the same gender, the DNA sequence is about 99.9% identical. Those 0.1% differences consist of DNA sequence variations that define our individuality. Single-nucleotide polymorphisms (SNPs) are the most common type of genetic variation in humans. Many of these are neutral polymorphisms while others represent disease-causing mutations. A major challenge is to identify SNPs and to learn how and why they cause disease and

TABLE 18-9 Gateways to Locus-Specific Databases

Site	Description	URL
GeneDis	From Tel Aviv University; performs pairwise alignments against a disease database	▸http://life2.tau.ac.il/GeneDis/
HUGO Mutation Database Initiative	A huge list of locus-specific mutation databases	▸http://ariel.ucs.unimelb.edu.au:80/~cotton/glsdb.htm
Human Gene Mutation Database	From the Institute of Medical Genetics in Cardiff	▸http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html
Universal Mutation Database	Software and databases for mutations in human genes, from INSERM	▸http://www.umd.necker.fr/
The Mammalian Gene Mutation Database (MGMD)	Database of published mutagen-induced gene mutations in mammalian tissues	▸http://lisntweb.swan.ac.uk/cmgt/index.htm
Sequence Variation Database (SVD) project	At EMBL-EBI	▸http://www2.ebi.ac.uk/mutations/
Gene-Specific Mutation Databases	Engelhardt Institute of Molecular Biology (Russia)	▸http://wgen.eimb.relarn.ru/databases/genespec.htm

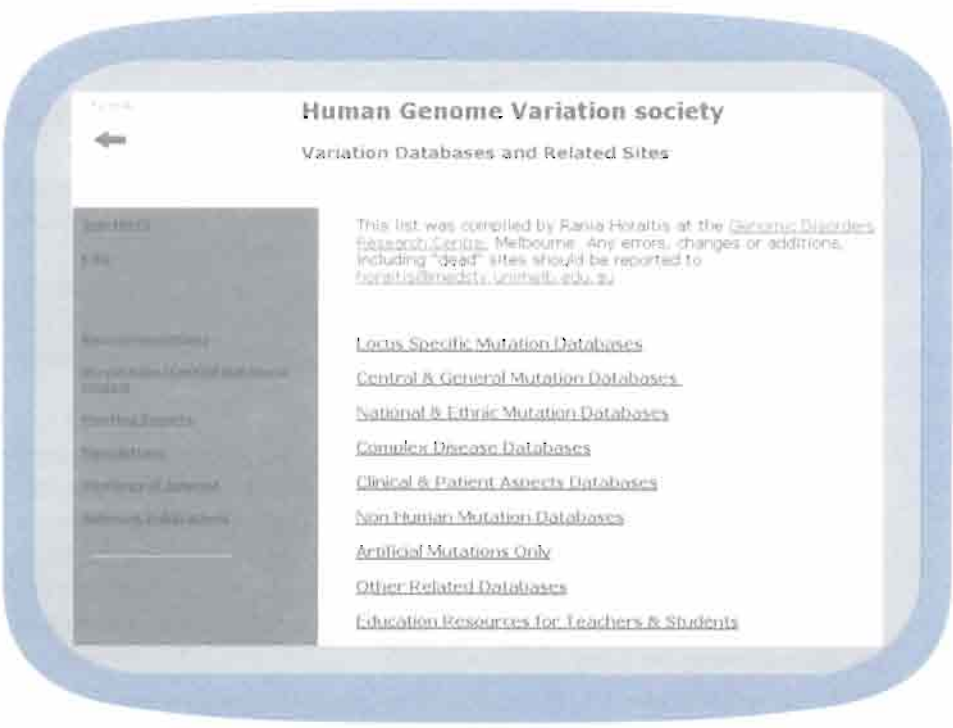


FIGURE 18.12. *The Human Genome Variation Society links to many gene-specific human mutation databases (http://ariel.ucs.unimelb.edu.au:80/~cotton/dblist.htm).*

Entry M	Short Citation	Nucleotide change	Amino acid change	Type of sequence change	Domain change location	Phenotype	Mutation/polymorphism	Sex
1271	Klauck, Pubmed [redacted]	c.419C>T	p.A140V	Missense	MBD	Not Rett synd. - Unaffected family member		F
1272	Klauck, Pubmed [redacted]	c.419C>T	p.A140V	Missense	MBD	Not Rett synd. - X-linked mental retardation, manic-depressive illness		M
1561	Carrozzi, ...	c.422_423insA	p.Y141fs	Frameshift insertion or deletion	MBD	Not known	Mutation associated with disease	F
330	De Bona, ... Pubmed [redacted]	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - Classical	Mutation associated with disease	F
482	Directly submitted	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - Classical	Mutation associated with disease	U
483	Directly submitted	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - Classical	Mutation associated with disease	U
1060	Shoyes, ... Pubmed [redacted]	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - Not certain	Mutation associated with disease	F
1154	Watson, ... Pubmed [redacted]	c.423C>G	p.Y141X	Nonsense	MBD	Not Rett synd. - Angelman syndrome	Mutation associated with disease	F
1434	Schwartzman, ... Pubmed [redacted]	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - Male variant	Mutation associated with disease	M
1461	Vacca, Pubmed [redacted]	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - Classical	Mutation associated with disease	F
1521	Yamada, ... Pubmed [redacted]	c.423C>G	p.Y141X	Nonsense	MBD	Rett syndrome - classical	Mutation associated with disease	U
45	Hoffbuhr, ... Pubmed [redacted]	c.426C>T	p.F142F	Silent	MBD	Rett syndrome - Not certain	Silent polymorphism	F

FIGURE 18.13. *RettBASE (http://mecp2.chw.edu.au/mecp2/) is an example of a locus-specific database. It describes nucleotide changes that have been reported at various positions of the* MECP2 *gene. Note that a substitution of C for G at position 423 causes a nonsense mutation (i.e., the introduction of a stop codon) in a site corresponding to the methyl-binding domain (MBD) of the protein. This mutation has been described in a number of patients, but the phenotypes include classical Rett syndrome in females and a different phenotype in a male patient.*

whether they are associated with disease-causing genes (Wang and Moult, 2001; Goodstadt and Ponting, 2001).

SNPs may account for 90% of the sequence differences between humans (Collins et al., 1998). You can read an overview of SNPs at the NCBI website (▸http://www.ncbi.nlm.nih.gov/SNP/get.html.cgi?whichHtml=overview).

The density of SNPs averages about 1 every 800 bp. SNPs can occur in the coding region of genes (causing synonymous or nonsynonymous changes), in introns, in regulatory regions of genes, or in intergenic regions.

An individual's SNP genotype may provide insight into his or her susceptibility to disease. This is because sequence variation may be associated with heritable phenotypes. There are four main ways that SNP data can be informative with respect to disease:

1. Functional Variation. A SNP may be associated with a nonsynonymous substitution in a coding region. In this case, the amino acid sequence of a protein is altered. There are an estimated 5.3 million SNPs in the human genome, each having a frequency of at least 10% (Kruglyak and Nickerson, 2001). According to this estimate, the frequency of SNPs is about one every 600 base pairs. (For a minimal allele frequency of 1%, there are an estimated 11 million SNPs, occurring one every 290 base pairs.)
2. Regulatory Variation. A SNP can occur in a noncoding region but influence the properties of gene expression (such as the time course, regional pattern, or physiological response) (Cowles et al., 2002).
3. Association. SNPs may be used in whole-genome association studies (Kruglyak, 1999). In linkage studies, genetic markers are used to search for coinheritance of chromosomal regions within families. In contrast, in association studies, unrelated affected individuals and control cases are studied to find differences in the frequency of genetic variants. SNPs represent an abundant type of sequence differences that can be used in association studies (Sherry et al., 2000). Thus, even if SNPs themselves do not cause changes that induce disease, they may be proximal to deleterious mutations.
4. Haplotype maps. While DNA variation occurs across all the chromosomes, some variations occur in well conserved regions called haplotype blocks. A haplotype is a combination of alleles that occur in a population; more generally, haplotypes define patterns of DNA sequence variation. Groups of adjacent SNPs on any given chromosome are inherited as sets or haplotype blocks. These blocks may contain few or many SNPs, and the size of a block may vary up to millions of base pairs. SNPs may be used as tags to define the behavior of entire haplotype blocks. Efforts are underway to define the structure of haplotype blocks in the human genome and other genomes (Patil et al., 2001; Gabriel et al., 2002; Phillips et al., 2003). Such a haplotype map (HapMap) can be used to select SNPs that are informative in explaining diversity among different ethnic groups and other populations. The HapMap project may also show which SNPs are most useful to examine the entire human genome for association with a disease phenotype.

The HapMap project is described at the National Human Genome Research Institute (NHGRI) website (▸http://www.genome.gov/page.cfm?pageID=10001688). The Haplotype Map project Data Coordinating Center website, located at Cold Spring Harbor Laboratory, is at ▸http://hapmap.cshl.org/.

In addition to their interest as *possible* links to disease-causing mutations, SNPs are also of interest by evolutionary biologists. Analyzing SNPs shared by ethnic groups may help trace human evolutionary history.

SNPs are discovered through several sources:

- EST databases can be mined for single-nucleotide differences (Buetow et al., 1999).
- Human genomic sequences can be analyzed to identify SNPs (Taillon-Miller et al., 1998).

TABLE 18-10 SNP Resources

Resource	Comment	URL
dbSNP	At NCBI	▸www.ncbi.nlm.nih.gov/SNP
Human SNP database	At the Whitehead Institute	▸http://www-genome.wi.mit.edu/snp/human/
The SNP Consortium (TSC)	A collaboration of industrial and academic laboratories	▸http://snp.cshl.org

Regardless of how SNPs are identified, it is important to determine whether they are authentic sequence differences or mere artifacts (such as sequencing errors from long sequencing reads or polymerase chain reaction amplification artifacts).

There are several major repositories of SNP data (Table 18.10). The dbSNP at NCBI is a repository of over 3 million human SNPs (Sherry et al., 2000). The database also includes microsatellites and small insertion/deletion polymorphisms. The current holdings of dbSNP are listed in Table 18.11.

We can explore dbSNP at NCBI by searching for *MECP2* variants. The output includes data on polymorphisms derived from mRNA alignments and from contig annotations (Fig. 18.14). SNPs are color coded in a graphical view and a tabular format to show coding region variations (either synonymous or nonsynonymous) and changes in untranslated regions, introns, and splice sites.

dbSNP was established in a collaboration between NCBI and the National Human Genome Research Institute. Currently, the density of SNPs in the human genome is about 8 per 10,000 bp. Note that the term "SNP" in dbSNP denotes a "variation" since the database contains variants other than SNPs.

To explore whether SNPs have been identified in a sequence of interest, you can use a dbSNP BLAST server at ▸http://www.ncbi.nlm.nih.gov/SNP/snpblastByChr.html.

COMPLEX DISORDERS

In contrast to single-gene disorders, complex disorders are very common in the population (Todd, 2001). These traits do not segregate in a simple, discrete, Mendelian manner. Examples are asthma, autism (Box 18.3), diabetes, high blood pressure, obesity, and osteoporosis. It is likely that the vast majority of human diseases involve multiple genes. Complex disorders are characterized by the following features:

- Multiple genes are involved. It is the combination of mutations in multiple genes that defines the disease. In single-gene disorders, even if there are modifying loci, one gene has a dramatic influence on the disease phenotype.

For a series of articles on complex disorders published in *Science*, see Kiberstis and Roberts (2002) as well as accompanying articles on diabetes, lupus, schizophrenia, and viewpoints on how to approach the study of complex disorders.

A quantitative trait locus (QTL) is an allele that contributes to a multifactorial disease.

TABLE 18-11 dbSNP Statistics (NCBI Genome Build 30, November 2002)

See ▸http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi. SNP count is the number of distinct RefSNPs having the noted functional relationship to at least one mRNA in the current assembly. Gene count is the number of distinct locus_id(s) having at least one variation of the noted functional class. (Genes with multiple variations may be counted in multiple classes.)

Functional Classification	SNP Count	Gene Count
Locus region	291,459	26,482
Allele synonymous to contig nucleotide	12,322	7,147
Allele nonsynonymous to contig nucleotide	16,251	8,496
Untranslated region	131,987	13,208
Intron	904,573	22,113
Splice site	277	268
Allele is same as contig nucleotide	28,491	11,621
Coding: synonymy unknown	13,501	3,584

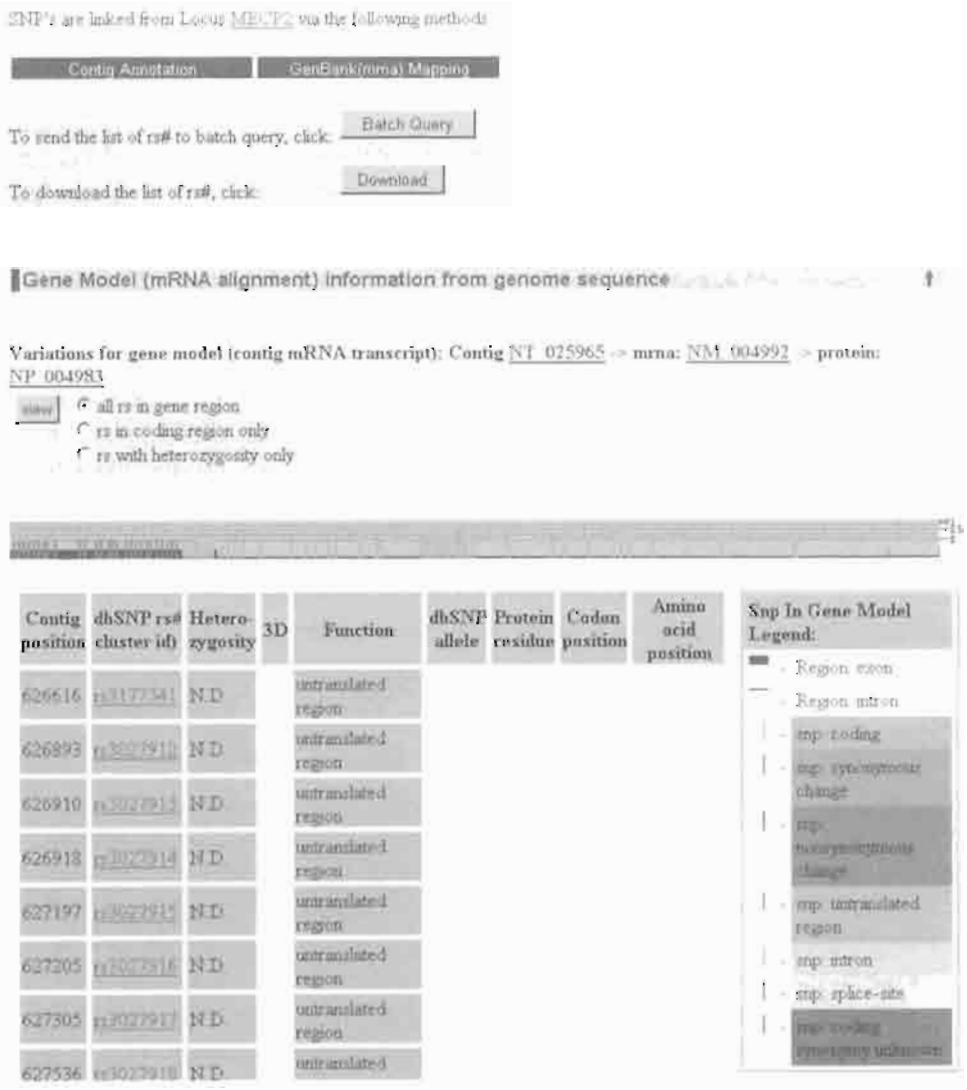


FIGURE 18.14. *The database of single-nucleotide polymorphisms (dbSNP) at NCBI shows SNPs associated with the* MECP2 *gene.*

- Complex diseases involve the combined effect of multiple genes, but they also are caused by both environmental factors and behaviors that elevate the risk of disease.
- Complex diseases are non-Mendelian: They show familial aggregation but not segregation. For example, autism is a highly heritable condition (if one identical twin is affected, there is a very high probability that the other is also affected).
- Susceptibility alleles have a high population frequency; that is, complex diseases are generally more frequent than single-gene disorders. Sickle cell anemia is unusually frequent in the African-American population, for a single-gene disorder, but the heterozygous condition confers a selective advantage (see Box 18.1 above).

BOX 18-3
Autism: Complex Disorder of Unknown Etiology

Autism (OMIM *209850) is a lifelong neurological disorder with onset before three years of age (Kanner, 1943; reviewed in Rapin and Katzman, 1998; Bailey et al., 1996; Ciaranello and Ciaranello, 1995; Piven, 1997; Rapin, 1997; Waterhouse et al., 1996; Gillberg and Coleman, 2000; Lord et al., 2000). It is characterized by a triad of deficits: (1) an individual's failure to have normal reciprocal social interaction, (2) impaired language or communication skills, and (3) restricted, stereotyped patterns of interests and activities. Autistic children's play is abnormal beginning in infancy, and there is a notable lack of imaginative play. Approximately 30% of autistic children appear to develop normally but then undergo a period of regression in language skills between 18 and 24 months of age. In addition, cognitive function may be impaired. Seventy-five percent of autistic individuals have mental retardation. Approximately 10% of autistic individuals have savant-like superior abilities in areas such as mathematical calculation, rote memory, or musical performance. Autism is accompanied by seizures; by adulthood about one-third of autistic individuals will have had at least two unprovoked seizures (Rossi et al., 1995; Olsson et al., 1988; Volkmar and Nelson, 1990).

The prevalence of autism has been estimated as between 0.2 and 2 per 1000 individuals (Rapin and Katzman, 1998; Fombonne, 1999; Smalley et al., 1988; Gillberg and Wing, 1999). About three to four times more males are affected than females (Fombonne, 1999).

The cause of autism is unknown, but there is strong evidence that the disorder is genetic (Smalley et al., 1988; Szatmari et al., 1998; Turner et al., 2000). The concordance between monozygotic twins is approximately 60%, and >90% if coaffected twins are defined as having classically defined autism or more generalized impairments in social skills, language, and cognition (Bailey et al., 1995). Autism has a far stronger genetic basis than most other common neuropsychiatric disorders such as schizophrenia or depression. Most genetic linkage studies have not resulted in the identification of genes that are significantly associated with autism, although the long arm of chromosome 15 has been implicated (Cook et al., 1997, 1998; Bundey et al., 1994; Flejter et al., 1996; Baker et al., 1994). Several genomewide scans with microsatellite markers failed to find significant linkage to a particular chromosome, although potential susceptibility regions were identified (International Molecular Genetic Study of Autism Consortium, 1998; Philippe et al., 1999; Lamb et al., 2000). A major conclusion of these studies is that autism is most likely caused by abnormalities in multiple, interacting genes (Gillberg, 1998). Known medical conditions affecting the central nervous system, such as fragile X syndrome and seizure disorder, may account for 10–30% of autistic cases (Barton and Volkmar, 1998).

- Susceptibility alleles have low penetrance. Penetrance is the frequency with which a dominant or homozygous recessive gene produces its characteristic phenotype in a population. At the extremes, it is an all-or-none phenomenon: A genotype is either expressed or it is not. In complex disorders, partial penetrance is common.

Penetrance is the frequency of manifestation of a hereditary condition in individuals. Having the genotype for a disease does not imply that the phenotype will occur, especially if multiple genes have modifying effects on the presentation of the phenotype.

How can we determine the causes of complex diseases? There are many approaches to finding treatments (or, ultimately, to finding cures): We will consider several of these, using the specific example of autism (Box 18.3):

- Heterogeneity is an issue for a variety of complex disorders. In the case of autism, clinical behavioral studies are performed to define the phenotypes of the disease. Some patients have profound mental retardation while others have a normal intelligence quotient. The extent to which language skills are either delayed or absent is highly variable. Thus, it is possible that the term "autism" refers to a cluster of phenotypically related disorders with overlapping diagnostic boundaries.
- Risk factors are studied based upon epidemiological studies.

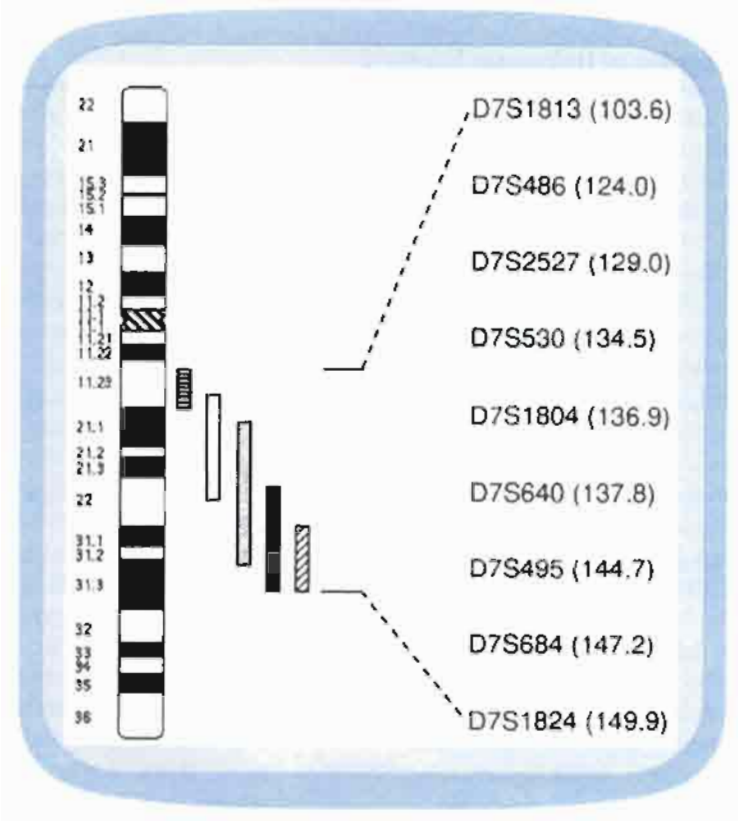


FIGURE 18.15. *Ideogram of chromosome 7 showing regions implicated in the etiology of autism based on linkage studies. Bioinformatics tools such as the NCBI Map Viewer are useful to define genes located in particular chromosomal regions. From Lamb et al. (2000). Used with permission.*

- The cause of autism may be sought through genetics studies, such as epidemiological measurements of the prevalence of the disorder in monozygotic and dizygotic twins. Twin studies and family studies reveal a strong genetic component to autism.
- Genetic defects may be sought through chromosomal analyses. Indeed, some patients with translocations have symptoms that resemble autism.
- Linkage and association analyses may be performed. Linkage studies of multiplex families have implicated chromosomes 7 and 15 and other loci as being involved in the etiology of autism. Figure 18.15 is an ideogram of chromosome 7 showing the regions implicated in several genomewide linkage studies (Lamb et al., 2000). The ideogram resembles that available in the NCBI Map Viewer (Fig. 17.2) and includes the Marshfield map showing marker distances from the p-arm telomere (in centimorgans).
- Pharmacological approaches can be used to discover drugs that improve the clinical symptoms of the disorder. Neuropathological studies can also define which regions of the brain are affected.
- Animal models can be developed, especially after a disease-causing gene has been identified.

A major approach to complex disorders is to discover the genes that cause the disease when mutated. A related approach is to define the consequence of the disease-causing mutation on the body. For example, gene expression profiling using microarrays or related techniques (Chapter 6) can be used to define the

TABLE 18-12 Frequency of Chromosomal Disorders among Liveborn Infants

Abnormalities	Disorder	Frequency
Autosomal	Trisomy 13	1 in 15,000
	Trisomy 18	1 in 5000
	Trisomy 21 (Down syndrome)	1 in 600
Sex chromosome	Klinefelter syndrome (47,XXY)	1 in 700 males
	XYY syndrome (47,XYY)	1 in 800 males
	Triple X syndrome (47,XXX)	1 in 1000 females
	Turner syndrome (45,X or 45X/46XX or 45X/46,XY or isochromosome Xq)	1 in 1500 females

Source: From Beaudet et al. (2001). Used with permission.

transcriptional response of the affected tissues to the disease. This may suggest avenues for pharmacological intervention or may lead to the identification of molecular markers for the disease (Colantuoni et al., 2000; Purcell et al., 2001). Similarly, proteomic approaches can be applied to affected tissues.

Analysis of Chromosomal Abnormalities in Disease

Large-scale chromosomal abnormalities are extremely common causes of disease in humans. These large-scale changes include aneuploidies such as trisomies and monosomies. Trisomies 13 (Patau syndrome), 18 (Edwards syndrome), and 21 (Down syndrome) are the only autosomal trisomies that are compatible with life (Table 18.12). (Of these, trisomies 13 and 18 are typically fatal in the first years of life.) Many developmental abnormalities such as Cri-du-chat syndrome, Angelman syndrome, Prader Willi syndrome, Smith-Magenis syndrome, and various forms of mental retardation result from the gain or loss of chromosomal regions. Cancers can also result from such changes. Lupski (1998) has defined genomic disorders as those changes in the structure of the genome that cause disease.

An aneuploidy is the condition of having an abnormal number of chromosomes.

Chromosomal disorders are an extremely common feature of normal human development. Humans have a very low fecundity even relative to other mammals, with perhaps 50% to 75% of all human conceptions resulting in miscarriage. This low fecundity is primarily due to the common occurrence of chromosomal abnormalities (Wells and Delhanty, 2000; Voullaire et al., 2000):

- A woman who has already had one child (and thus is of established fertility) has only a 25% chance of achieving a viable pregnancy in any given menstrual cycle.
- 52% of all women that conceive have an early miscarriage.
- Following in vitro fertilization, pregnancies that are confirmed positive in the first two weeks result in miscarriage 30% of the time.
- Over 60% of spontaneous abortions that occur at 12 weeks gestation or earlier are aneuploid, suggesting that early pregnancy failures are likely due to lethal chromosome abnormalities.

The most common chromosomal aberrations in early development likely involve the gain or loss of whole chromosomes. Such structural abnormalities may

BOX 18-4
Genomic Microarrays

A recent approach to the detection of chromosomal abnormalities is the use of genomic microarrays (Hermsen et al., 2001; Pollack et al., 1999; Pinkel et al., 1998; Lichter et al., 2000). This approach has been called comparative genomic hybridization. In one variant, bacterial artificial chromosome (BAC) clones, corresponding to known chromosomal loci, are deposited on the surface of a glass slide. Each BAC is typically 200 kilobases (kb) in length, and the BAC clones may be spaced an average of 0.5–10 megabases (Mb) apart across the genome. The technique offers increasing resolution as the BAC clones on the array correspond to more densely spaced regions across the chromosomes.

Genomic DNA is purified from a patient and an appropriate control case, labeled with green or red fluorophores (e.g. Cy3 or Cy5), and cohybridized on an array. The ratio of the dyes reflects the relative amount of genomic DNA hybridizing to each BAC, and in most instances the ratio is 1:1. However, duplications or deletions can be identified by deviations from this ratio.

Genomic arrays resemble DNA microarrays that are used for gene expression analysis (Chapters 6 and 7). In each case, DNA is immobilized on a glass slide or other solid support. And in each case, a control and an experimental sample may be labeled with green and red dyes and cohybridized on the surface of the array. Following washing of the slide, signal is detected with a scanner. The main difference is that the for genomic arrays the sample consists of fragmented genomic DNA (rather than RNA or mRNA), and the information one obtains concerns the gain or loss of large regions of chromosomal DNA (e.g. several megabases) rather than information on the expression levels of individual genes.

be detected by standard cytogenetic approaches such as karyotype analysis and fluorescence in situ hybridization (FISH). These techniques may also reveal commonly observed phenomena such as large-scale duplications, deletions, or rearrangements involving many millions of base pairs.

A novel approach to the study of chromosomal abnormalities that occur during normal development is through the use of genomic microarrays (Box 18.4). Wells and Delhanty (2000) amplified genomic DNA from 12 normally developing human embryos that were obtained three days following in vitro fertilization. They analyzed DNA from 64 individual cells obtained from these embryos. By comparing the genomic DNA of each cell to known normal controls, they were able to measure the ratio of chromosomal copy number in each sample across the length of all chromosomes, at 10 megabase intervals. Nine of the 12 apparently normal embryos contained aneuploid cells. The abnormalities included autosomal monosomies and even three cases of nullisomy (total loss of both copies of a chromosome). Voullaire et al. (2000) applied the same approach independently. They amplified genomic DNA from 65 cells derived from nine normal embryos. Only three embryos showed a consistently normal chromosomal profile. They observed aneuploidy in the other cases, including examples of monosomy, trisomy, mosaicism, and chromosomal breakage. In both studies, examples of "chaotic" embryos were reported in which every cell that was examined exhibited extensive aneuploidy due to apparently random chromosomal segregation.

Together, these studies suggest that the majority of apparently normal human post-fertilization embryos actually have aneuploidies. Almost all these abnormalities will result in early embryonic lethality. These chromosomal aberrations account for the relatively low success rate of in vitro fertilization. Bioinformatics tools such as genomic microarrays provide a genome-wide view of the remarkable complexity of development at the level of chromosomal integrity. This technology may be used to understand chromosomal changes as a function of maternal or paternal age, or in a variety of species (including those with relatively high versus low fecundity).

TABLE 18-13 Web Resources for Study of Human Disease at Systems Level

Resource	Comment	URL
LensGDDB	Human Lens Genetic Disease Database	▸http://ken.mitton.com/ern/lensbase.html
Retina International	Four databases: protein, mutation, disease, and animal model	▸http://www.retina-international.org/sci-news/database.htm
National Cardiovascular Disease Database	Data on risk factors, death rates, literature	▸http://www.aihw.gov.au/cvdhtml/cvd-menu.htm
Gastrointestinal (GI) Disease Database	Resource gateway for dozens of GI diseases	▸http://www.gastro.net.au/gastrodiseases/
Muscular dystrophies	Information on genes, diseases, mutation databases, and resources	▸http://www.dmd.nl/
MITOMAP	A human mitochondrial genome database	▸http://www.mitomap.org/
Peroxisome website	Sections for physicians, scientists, and laypersons	▸http://www.peroxisome.org/

Systems-Level Bioinformatics Resources for Human Disease: Organellar and Pathway Databases

Eukaryotic cells are organized into organelles, such as the nucleus, endoplasmic reticulum, Golgi complex, peroxisome, and mitochondrion. Each organelle serves a specialized function, gathering particular protein products to form enzymatic reactions necessary for cell survival, separating metabolic processes, and segregating harmful products. We have considered human disease from the perspective of genes and gene products. We will next examine disease in the context of the higher organizational level of organelles and pathways. A variety of web resources are presented in Table 18.13.

Let us consider the mitochondrion. This organelle was described as the site of respiration in the 1940s, and mitochondrial DNA was first reported by Nass and Nass (1963). But it was not until 1988 that the first disease-causing mutations in mitochondria were described (Wallace et al., 1988; Holt et al., 1988). Today, over 100 disease-causing point mutations have been described (reviewed in DiMauro and Schon, 2001; Schon, 2000). The mitochondrial genome encodes 37 genes, any of which can be associated with disease (see Figs. 12.7 and 12.8). Figure 18.16 shows a morbidity map of the human mitochondrial genome.

Most mitochondrial proteins are the product of nuclear genes, and most mitochondrial diseases are caused by mutations in nuclear genes.

Mitochondrial genetics differs from Mendelian genetics in three main ways (DiMauro and Schon, 2001):

1. Mitochondrial DNA is maternally inherited. Mitochondria in the embryo are derived primarily from the ovum, while sperm mitochondria fail to enter the egg and are actively degraded. Thus, a woman having a mitochondrial DNA mutation may transmit it to her children, but only her daughters will further transmit the mutation to their children.
2. While nuclear genes exist with two alleles (one maternal and one paternal), mitochondrial genes exist in hundreds or thousands of copies per cell. (A typical mitochondrion contains about ten copies of the mitochondrial genome.) An individual may harbor varying ratios of normal and mutated

Normally all mitochondrial genomes are the same, a condition called homoplasy. Pathogenic mutations may be heteroplasmic (having a mixture of normal and mutated genomes).

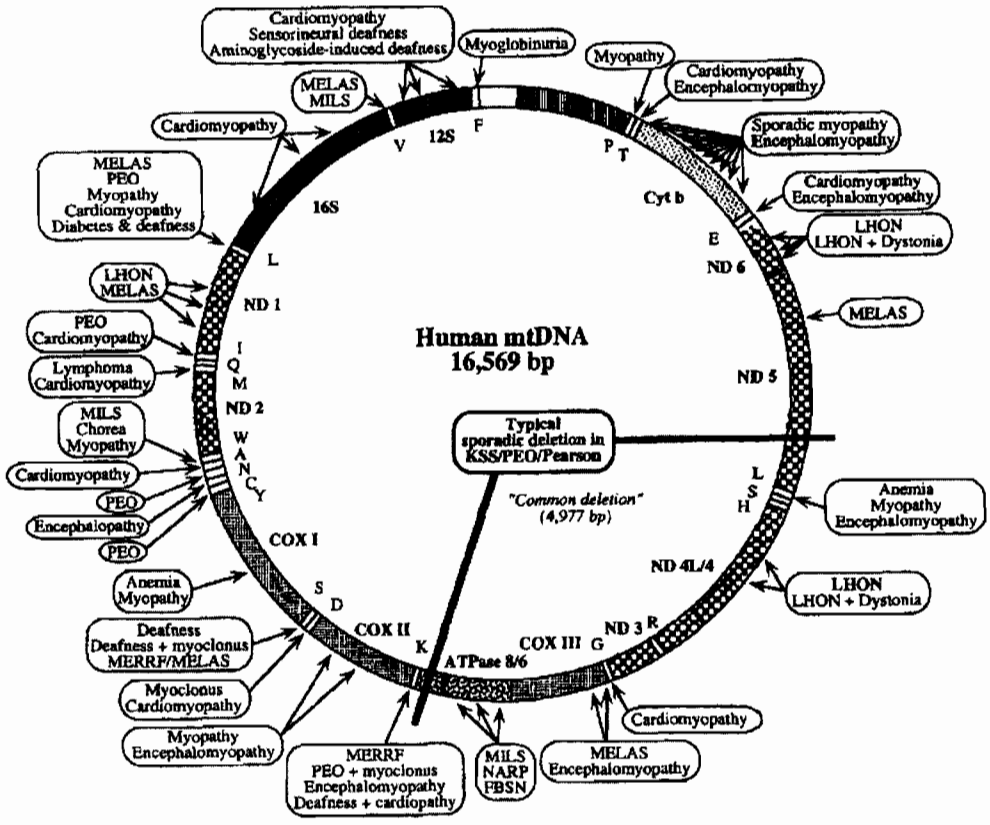


FIGURE 18.16. *Morbidity map of the human mitochondrial genome. Abbreviations are for the genes encoding seven subunits of complex I (ND), three subunits of cytochrome c oxidase (COX), cytochrome b (Cyt b), and the two subunits of ATP synthase (ATPase 6 and 8). 12S and 16S refer to ribosomal RNAs; 22 transfer RNAs are identified by the one-letter codes for the corresponding amino acids. FBSN, familial bilateral striatal necrosis; KSS, Kearns–Sayre syndrome; LHON, Leber hereditary optic neuropathy; MELAS, mitochondrial encephalomyopathy, lactic acidosis, and strokelike episodes; MERRF, myoclonic epilepsy with ragged-red fibers; MILS, maternally inherited Leigh syndrome; NARP, neuropathy, ataxia, retinitis pigmentosa; PEO, progressive external ophthalmoplegia. From DiMauro and Schon (2001). Used with permission.*

mitochondrial genomes. Some critical threshold of mutated mitochondrial genomes is required before a disease is manifested.

3. As cells divide, the proportion of mitochondria having mutated genomes can change, thus affecting the phenotypic expression of mitochondrial disorders. Clinically, mitochondrial disorders vary at different times and in different regions of the body. An extremely broad variety of diseases are associated with mutations in mitochondrial DNA.

MITOMAP is a comprehensive mitochondrial genome database (see Table 18.13 for URL). The site lists a broad variety of information on mutations and polymorphisms in mitochondrial genomes involving all known genetic mechanisms (inversions, insertions, deletions, etc.).

HUMAN DISEASE GENES IN MODEL ORGANISMS

The study of human disease genes and gene products in other organisms is of fundamental importance in our efforts to understand the pathophysiology of human

disease. While mutations in genes cause many diseases, it is the aberrant protein product that has the proximal functional consequence on the cell and ultimately on the organism. Once a human disease gene is identified in a model organism, it can often be knocked out or otherwise manipulated. This allows the phenotypic consequences of specific mutations to be assessed.

An important resource for the study of diseases in animals is Online Mendelian Inheritance in Animals (OMIA) (▸http://www.angis.org.au/Databases/BIRX/omia/).

A basic question then is to identify which known human disease genes have orthologs in model organisms. This approach is of interest even though the consequence of mutating that ortholog may differ. A group of 55 authors collaborated on a systematic sequence analysis of the *Drosophila melanogaster, Caenorhabditis elegans,* and *Saccharomyces cerevisiae* genomes (Rubin et al., 2000). They identified 289 genes that are mutated, altered, amplified, or deleted in human disease. Of these genes, 177 (61%) were found to have an ortholog in *Drosophila.* These data are displayed in Figure 18.17, showing the presence of fly, worm, and yeast orthologs to human disease genes that are functionally categorized in cancer, neurological, cardiovascular, endocrine, and other disease types. Reiter et al. (2001) extended this study to 929 human disease genes in OMIM, 714 of which (77%) matched 548 *Drosophila* protein sequences (Table 18.14). The Reiter et al. (2001) data have been deposited in Homophila, a *Homo sapiens/Drosphila* disease database.

At the time that *C. elegans* was sequenced, about 65% of human disease genes had identifiable *C. elegans* orthologs (Ahringer, 1997).

The Homophila website is at ▸http://homophila.sdsc.edu/.

The cataloguing of human disease genes in model organisms is important in our efforts to establish functional assays for these genes. In addition to the results in *S. cerevisiae, D. melanogaster,* and *C. elegans,* similar descriptions have been made in other eukaryotes, such as *Schizosaccharomyces pombe* (Wood et al., 2002) and *Arabidopsis* (The *Arabidopsis* Genome Initiative, 2000). For *S. pombe,* orthologs were identified both for human cancer genes (Table 18.15) and a variety of neurological, metabolic, and other disorders (Table 18.16).

It is perhaps expected that human genes involved in cancer are also present in fungi; examples include genes encoding proteins involved in DNA damage and repair and the cell cycle. It might seem surprising that genes implicated in neurological disorders are present in single-celled fungi. However, the explanation may be that neurons are a particularly susceptible cell type with unique metabolic requirements. For example, most lysosomal disorders are caused by the loss of an enzyme that normally contributes to lysosmal function or to intracellular trafficking to lysosomes. Multiple organ systems are typically compromised, but neurological features such as mental retardation are a common consequence of these disorders. The lysosome is a primary site for catabolism in the cell. In fungi, the vacuole performs similar functions, and many human homologs of fungal vacuolar proteins have been identified.

What is the significance of having identified human disease gene homologs? Beyond cataloguing the presence of orthologs, a next step is to relate the information on mutations in human disease to the conservation of amino acid residues in orthologs. Miller and Kumar (2001) selected seven genes that when mutated cause disease in humans: the cystic fibrosis transmembrane regulator (*CFTR*), glucose-6-phosphate dehydrogenase (*G6PD*), neural cell adhesion molecule L1 (*L1CAM*), phenylalanine hydroxylase (*PAH*), paired box 6 (*PAX6*), the X-linked retinoschisis gene (RS1), and a tuberous sclerosis gene (*TSC2*). For each of these genes, two resources are available: locus-specific databases of mutations that occur in patients and the sequences of a variety of metazoan homologs (e.g., primates, rodents, fish, insects, nematodes). They generated multiple sequence alignments for each of these seven genes to test the null hypothesis that point mutations occur randomly throughout each gene. Following statistical tets (χ^2 analysis), they determined that

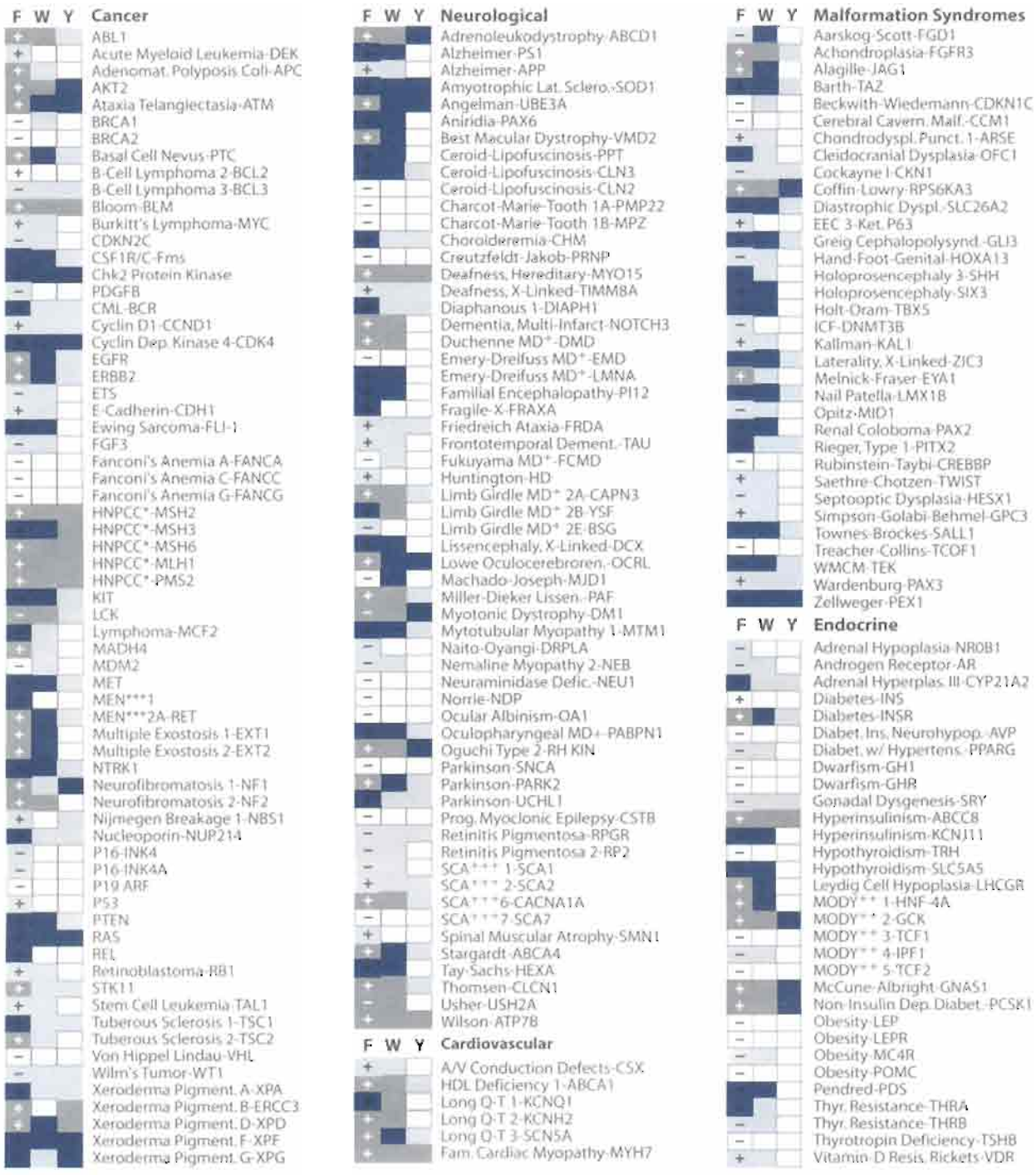


FIGURE 18.17. *A set of 289 proteins encoded by human disease genes were used as blastp queries against a set of 38,860 proteins from the complete genomes of a fly (F), a yeast (Y), and a worm (W). Database matches are presented according to their level of statistical significance. White boxes represent* E *values greater than* 1×10^{-6} *(no or weak similarity). Light gray boxes represent* E *values from* 1×10^{-6} *to* 1×10^{-40}. *Red boxes represent* E *values from* 1×10^{-40} *to* 1×10^{-100}. *Dark gray boxes represent* E *values below* 1×10^{-100}. *A plus sign indicates that the* Drosophila *protein is the functional equivalent of the human protein (based on criteria including sequence similarity, InterPro domain composition, and supporting biological evidence). A minus sign indicates that evidence was not obtained for functional equivalence to the human protein. Adapted from Rubin et al. (2000). Used with permission.*

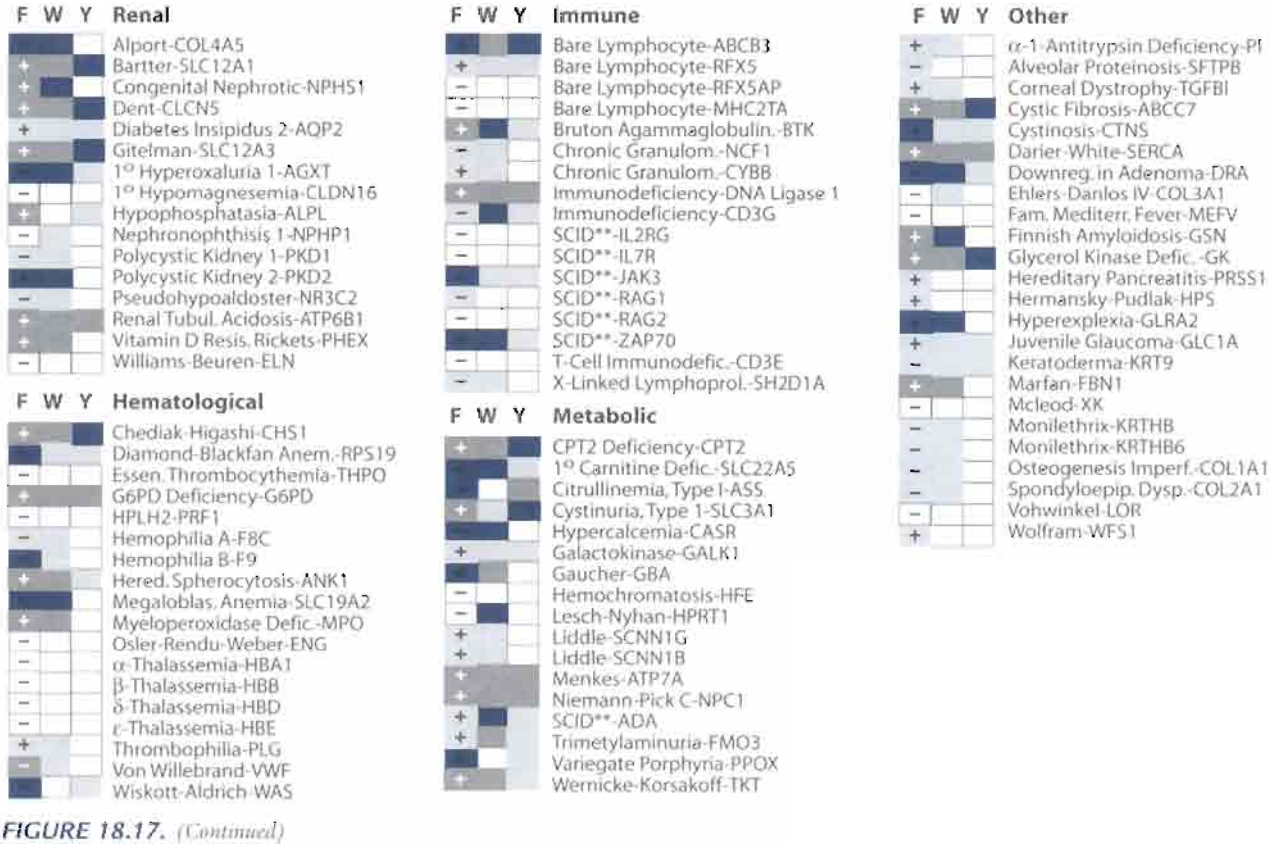


FIGURE 18.17. (*Continued*)

most amino acids that can produce human disease mutations are conserved (at least among mammals). Variable sites correspond to positions where amino acid changes are tolerated due to relaxed selection constraints.

As we discussed in Chapter 3, PAM and BLOSUM matrices (Figs. 3.10–3.17) reveal that the relative rates of evolutionary substitution vary for different pairs of amino acids. Glutamic acid commonly changes to aspartic acid (in the PAM250 matrix, the score is +2); these two residues are both acidic and thus share common physiochemical properties. However, glutamic acid rarely changes to lysine (the PAM250 score is 0). In human disease, a glutamic acid-to-lysine mutation commonly causes disease. Miller and Kumar (2001) displayed these findings in a table showing the relative frequencies of amino acid changes observed in a variety of eukaryotes (Fig. 18.18, circles) versus amino acid changes that have been detected in patients (Fig. 18.18, squares).

These analyses suggest that disease-associated changes tend to occur at conserved residues. Furthermore, the amino acid changes found in human disease do not commonly occur in comparisons between species.

In Mouse

The mouse genome, reported by the Mouse Genome Sequencing Consortium (Waterston et al., 2002), presents us with the most important animal model of

TABLE 18-14 Classification of 714 *Drosophila* Genes According to Human Disease Phenotypes

Disorder	Number Genes
Neurological	
Neuromuscular	20
Neuropsychiatric	9
CNS/developmental	8
CNS/ataxia	9
Mental retardation	6
Other	22
Total	74
Endocrine	
Diabetes	10
Other	40
Total	50
Deafness	
Syndromic	7
Nonsyndromic	6
Total	13
Cardiovascular	
Cardiomyopathy	10
Conduction defects	4
Hypertension	7
Atherosclerosis	3
Vascular malformations	2
Total	26
Ophthalmological	
Anterior segment	
Aniridia	1
Rieger syndrome	1
Mesenchymal dysgenesis	2
Iridogoniodysgenesis	2
Corneal dystrophy	2
Cataract	3
Glaucoma	2
Subtotal	13
Retina	
Retinal dystrophy	1
Choroideremia	1
Color vision defects	4
Cone dystrophy	2
Cone rod dystrophy	1
Night blindness	8
Leber congenital amaurosis	2
Macular dystrophy	4
Retinitis pigmentosa	7
Subtotal	30
Total	43
Pulmonary	4
Gastrointestinal	13
Renal	13
Immunological	
Complement mediated	11
Other	22
Total	33
Hematological	
Erythrocyte, general	29
Porphyrias	7
Platelets	6
Total	42
Coagulation abnormalities	28
Malignancies	
Brain	3
Breast	4
Colon	11
Other gastrointestinal	3
Genitourinary	5
Gynocological	3
Endocrine	3
Dermatological	3
Xeroderma pigmentosa	6
Other/sarcomas	9
Hematological malignancies	29
Total	79
Skeletal development	
Craniosynostosis	5
Skeletal dysplasia	13
Other	8
Total	26
Soft tissue	2
Connective tissue	18
Dermatological	25
Metabolic/mitochondrial	123
Pharmacological	12
Peroxisomal	9
Storage	
Glycogen storage	11
Lipid storage	13
Mucopolysaccharidosis	10
Other	3
Total	37
Pleiotropic developmental	
Growth, immune, cancer	7
Apoptosis	1
Other	27
Total	35
Complex other	9
Total	714

Source: Adapted from Reiter et al. (2001). Used with permission.

TABLE 18-15 ***Schizosaccharomyces pombe* Genes Related to Human Cancer Genes**

Score is the expect value from a BLAST search; a score of $<1 \times 10^{-40}$ refers to a score between $<1 \times 10^{-40}$ and 1×10^{-100}

Human Cancer Gene	Score	*S. pombe* Gene/Product	Systematic Name
Xeroderma pigmentosum D; *XPD*	$<1 \times 10^{-100}$	rad15, rhp3	SPAC1D4.12
Xeroderma pigmentosum B; *ERCC3*	$<1 \times 10^{-100}$	rad25	SPAC17A5.06
Hereditary nonpolyposis colorectal cancer (HNPCC); *MSH2*	$<1 \times 10^{-100}$	rad16, rad10, rad20, swi9	SPBC24C6.12C
Xeroderma pigmentosum F; *XPF*	$<1 \times 10^{-100}$	cdc17	SPCC970.01
HNPCC; *PMS2*	$<1 \times 10^{-100}$	pms1	SPAC57A10.13C
HNPCC; *MSH6*	$<1 \times 10^{-100}$	msh6	SPAC19G12.02C
HNPCC; *MSH3*	$<1 \times 10^{-100}$	swi4	SPCC285.16C
HNPCC; *MLH1*	$<1 \times 10^{-100}$	mlh1	SPAC8F11.03
Haematological Chediak–Higashi syndrome; *CHS1*	$<1 \times 10^{-100}$	–	SPBC1703.4
Darier–White disease; *SERCA*	$<1 \times 10^{-100}$	Pgak	SPBC28E12.06C
Bloom syndrome; *BLM*	$<1 \times 10^{-100}$	Hus2, rqh1, rad12	SPBC31E1.02C
Ataxia telangiectasia; *ATM*	$<1 \times 10^{-100}$	Tel1	SPAC2G11.12
Xeroderma pigmentosum G; *XPG*	$<1 \times 10^{-40}$	rad13	SPBC3E7.08C
Tuberous sclerosis 2; *TSC2*	$<1 \times 10^{-40}$	–	SPAC630.13C
Immune bare lymphocyte; *ABCB3*	$<1 \times 10^{-40}$	–	SPBC9B6.09C
Downregulated in adenoma; *DRA*	$<1 \times 10^{-40}$	–	SPAC869.05C
Diamond–Blackfan anemia; *RPS19*	$<1 \times 10^{-40}$	rps19	SPBC649.02
Cockayne syndrome 1; *CKN1*	$<1 \times 10^{-40}$	–	SPBC577.09
RAS	$<1 \times 10^{-40}$	Ste5, ras1	SPAC17H9.09C
Cyclin-dependent kinase 4; *CDK4*	$<1 \times 10^{-40}$	Cdc2	SPBC11B10.09
CHK2 protein kinase	$<1 \times 10^{-40}$	Cds1	SPCC18B5.11C
AKT2	$<1 \times 10^{-40}$	Pck2, sts6, pkc1	SPBC12D12.04C

Source: Adapted from Wood et al. (2002). Used with permission.

human disease. A number of important resources are available:

- The FANTOM database, part of the RIKEN Mouse Gene Encyclopedia Project, contains information on full-length mouse cDNA clones (Bono et al., 2002). The associated online database includes a description of full-length clones corresponding to human disease genes (Table 18.17).

TABLE 18-16 ***Schizosaccharomyces pombe*** **Genes Related to Human Disease Genes**

Score is the expect value from a BLAST search

Human Cancer Gene	Disease	Score	*S. pombe* Gene/Product
Wilson disease; *ATP7B*	Metabolic	$<1 \times 10^{-100}$	P-type copper ATPase
Non-insulin-dependent diabetes; *PCSK1*	Metabolic	$<1 \times 10^{-100}$	Krp1, kinesin related
Hyperinsulinism; *ABCC8*	Metabolic	$<1 \times 10^{-100}$	ABC transporter
G6PD deficiency; *G6PD*	Metabolic	$<1 \times 10^{-100}$	Zwf1 GP6 dehydrogenase
Citrullinemia type I; *ASS*	Metabolic	$<1 \times 10^{-100}$	Arginosuccinate synthase
Wernicke–Korsakoff syndrome; *TKT*	Metabolic	$<1 \times 10^{-40}$	Transketolase
Variegate pophyria; *PPOX*	Metabolic	$<1 \times 10^{-40}$	Protoporphyrinogen oxidase
Maturity-onset diabetes of the young (MODY2); *GCK*	Metabolic	$<1 \times 10^{-40}$	Hxk1, hexokinase
Gitelman's syndrome; *SLC12A3*	Metabolic	$<1 \times 10^{-40}$	CCC Na-K-Cl transporter
Cystinuria type 1; *SLC3A1*	Metabolic	$<1 \times 10^{-40}$	α-Glucosidase
Cystic fibrosis; *ABCC7*	Metabolic	$<1 \times 10^{-40}$	ABC transporter
Bartter's syndrome; *SLC12A1*	Metabolic	$<1 \times 10^{-40}$	CCC Na-K-Cl transporter
Menkes syndrome; *ATP7A*	Neurological	$<1 \times 10^{-100}$	P-type copper ATPase
Deafness, hereditary; *MYO15*	Neurological	$<1 \times 10^{-100}$	Myo51 class V myosin
Zellweger syndrome; *PEX1*	Neurological	$<1 \times 10^{-40}$	AAA-family ATPase
Thomsen disease; *CLCN1*	Neurological	$<1 \times 10^{-40}$	ClC chloride channel
Spinocerebellar ataxia type 6 (SCA6); *CACNA1A*	Neurological	$<1 \times 10^{-40}$	VIC sodium channel
Myotonic dystrophy; *DM1*	Neurological	$<1 \times 10^{-40}$	Orb6 Ser/Thr protein kinase
McCune–Albright syndrome; *GNAS1*	Neurological	$<1 \times 10^{-40}$	Gpa1 GNP
Lowe's oculocerebrorenal syndrome; *OCRL*	Neurological	$<1 \times 10^{-40}$	PIP phosphatase
Dents; *CLCN5*	Neurological	$<1 \times 10^{-40}$	ClC chloride channel
Coffin–Lowry; *RPS6KA3*	Neurological	$<1 \times 10^{-40}$	Ser/Thr protein kinase
Angelman; *UBE3A*	Neurological	$<1 \times 10^{-40}$	Ubiquitin–protein lgase
Amyotrophic lateral sclerosis; *SOD1*	Neurological	$<1 \times 10^{-40}$	Sod1, superoxide dismutase
Oguschi type 2; *RHKIN*	Neurological	$<1 \times 10^{-40}$	Ser/Thr protein kinase
Familial cardiac myopathy; *MYH7*	Cardiac	$<1 \times 10^{-100}$	Myo2, myosin II
Renal tubular acidosis; *ATP6B1*	Renal	$<1 \times 10^{-100}$	V-type ATPase

Source: Adapted from Wood et al. (2002). Used with permission.
Abbreviation: GNP, guenine nucleotide binding.

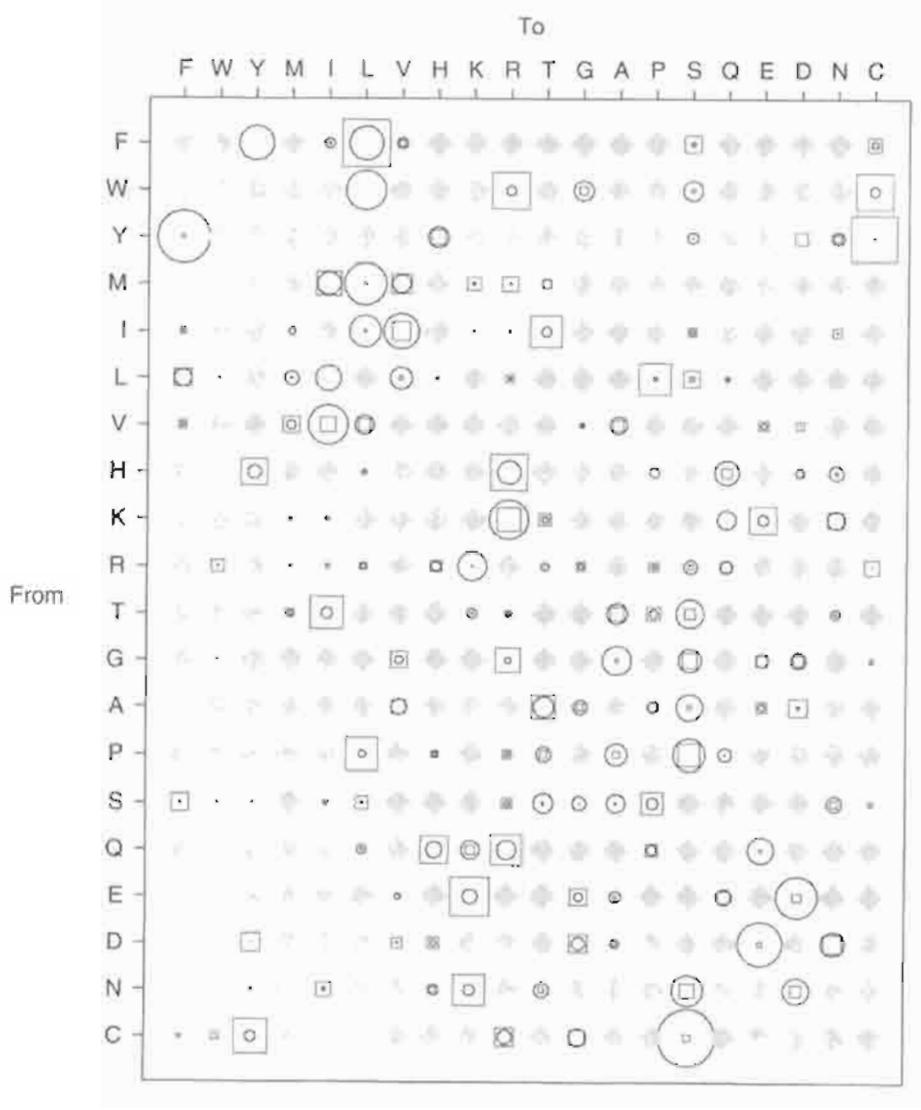


FIGURE 18.18. *Amino acid substitutions that occur in human disease are generally not allowed by natural selection. The figure shows a table of the amino acids, indicating the relative frequencies of amino acid changes observed in comparisons between various eukaryotic species (circles) and those changes detected in patients with diseases (squares). The size of the symbols is proportional to the relative frequency of change for a given amino acid. Diamonds indicate changes that cannot be observed as a result of a single base mutation. From Miller and Kumar (2001). Used with permission.*

- The Jackson Laboratory website offers a list of mouse/human gene homologs, including mouse models for human disease.
- High-efficiency mutagens such as *N*-ethyl-*N*-nitrosourea (ENU) or radiation have been applied to mice to generate models of human disease (Hrabe de Angelis and Strivens, 2001). Nolan et al. (2002) discuss this approach in detail, including web resources for high-throughput screening centers and strategies for finding gene mutations that correspond to novel phenotypes.
- The Dysmorphic Human-Mouse Homology Database includes a description of these disorders and a human/mouse syntenic map.
- The Whole Mouse Catalog describes mouse models of human disease.

The Jackson Laboratory website "Mouse Models for Human Disease: Mouse/Human Gene Homologs" is available online (▸http://jaxmice.jax.org/jaxmicedb/html/model.975.shtml)

This database, offered by the University of London, is at ▸http://www.hgmp.mrc.ac.uk/DHMHD/dysmorph.html.

You can access this mouse information at ▸http://www.rodentia.com/wmc/domain_genome.html#transgenics and ▸http://www.rodentia.com/wmc/domain_mouse.html.

The sequencing of the mouse genome was achieved by both Celera Genomics and by a public consortium (Chapter 16). Celera sequenced the genomic DNA of several mouse strains and noted their differences in susceptibility to infectious disease (Table 18.18) and complex inherited disease (Table 18.19). Comparative

TABLE 18-17 Full-Length Mouse cDNA Clones Corresponding to Human Disease Genes in FANTOM2 Database

See▸ http://fantom2.gsc.riken.go.jp/supplement/disease_genes/. The table at this URL is clickable, leading to a database of gene identifiers with links to OMIM and LocusLink

		FANTOM2 clones	
		Yes	No
Mouse ortholog	Yes	740	173
	No	67	13

genomic data will likely help explain why some mouse strains vary in their disease susceptibility. Adrian Hill (2001) has reviewed the genomics and genetics of infectious disease susceptiblility in humans.

The public consortium that sequenced the mouse genome reported that 687 human disease genes have clear orthologs in mouse (Waterston et al., 2002). Surprisingly, for several dozen genes, the wild-type mouse gene sequence was identical to the sequence that is associated with disease in humans. These genes are listed in Table 18.20. This suggests that, assuming the mouse does not have these diseases, any mouse model for these diseases must be used with caution. Conceivably, mice have modifying genes (or paralogous genes) not present in humans. Also, inbred strains of laboratory mice are exposed to different environmental stressors than mice in the wild, and their disease susceptibility could vary.

HUMAN DISEASE ORGANIZATIONS

We have defined bioinformatics as the use of computer algorithms and computer databases to study genes, proteins, and genomes. For human disease, a number of databases are available on the World Wide Web (Guttmacher, 2001). Table 18.21 lists some of these resources, including organizations that provide information to families of those with any of several hundred different diseases.

FUNCTIONAL CLASSIFICATION OF DISEASE GENES

We conclude our study of human disease by considering the principles of human disease. The variety of human diseases is extraordinarily broad, yet the field of bioinformatics may provide insight into a logic of disease. One such attempt was by

TABLE 18-18 Infectious Disease Susceptibility of Mouse Strains

	Inbred Mouse Strain	
Infectious Disease	A/J	C57BL/6J
Legionnaire's pneumonia	Susceptible	Resistant
Malaria	Susceptible	Resistant
Viral (MHV3) hepatitis	Resistant	Susceptible
Murine AIDS	Resistant	Susceptible

Source: Adapted from▸ http://www.celera.com/genomics/other/mousec16pres/home.cfm?ppage=5&returnLoc=index.jsp.

TABLE 18-19 Common Complex Disease Susceptibility of Mouse Strains

	Inbred Mouse Strain	
Complex Disease	A/J	C57BL/6J
Arthritis	Susceptible	Resistant
Colon cancer	Susceptible	Resistant
Lung cancer	Susceptible	Resistant
Asthma	Susceptible	Resistant
Atherosclerosis	Resistant	Susceptible
Hypertension	Resistant	Susceptible
Type II diabetes	Resistant	Susceptible
Osteoporosis	Susceptible	Resistant
obesity	Resistant	Susceptible

Source: Adapted from http://www.celera.com/genomics/other/mousec16pres/home.cfm?ppage=5&returnLoc=index.jsp.

TABLE 18-20 Human Disease-Associated Sequence Variants for Which Wild-Type Mouse Sequence Matches Diseased Human Sequence

Disease	OMIM	Mutation
Hirschsprung disease	142623	E251K
Leukencephaly with vanishing white matter	603896	R113H
Mucopolysaccharidosis type IVA	253000	R376Q
Breast cancer	113705	L892S
	600185	V211A, Q2421H
Parkinson's disease	601508	A53T
Tuberous sclerosis	605284	Q654E
Bardet–Biedl syndrome, type 6	209900	T57A
Mesothelioma	156240	N93S
Long QT syndrome 5	176261	V109I
Cystic fibrosis	602421	F87L, V754M
Porphyria variegata	176200	Q127H
Non-Hodgkin's lymphoma	605027	A25T, P183L
Severe combined immunodeficiency disease	102700	R142Q
Limb-girdle muscular dystrophy type 2D	254110	P30L
Long-chain acyl-CoA dehydrogenase deficiency	201460	Q333K
Usher syndrome type 1B	276902	G955S
Chronic nonspherocytic haemolytic anemia	206400	A295V
Mantle cell lymphoma	208900	N750K
Becker muscular dystrophy	300377	H2921R
Complete androgen insensitivity syndrome	300068	G491S
Prostate cancer	176807	P269S, S647N
Crohn's disease	266600	W157R

Source: Adapted from Waterston et al. (2002). Used with permission.

TABLE 18-21 General Web Resources for Study of Human Diseases

Site	Description	URL
Diseases, Disorders and Related Topics	Karolinska Institute (Stockholm)	▸http://www.mic.ki.se/Diseases/
The Frequency of Inherited Disorders Database (FIDD)	From the Institute of Medical Genetics University of Wales College of Medicine	▸http://archive.uwcm.ac.uk/uwcm/mg/fidd/
GeneCards	A database of human genes, their products, and their involvement in diseases	▸http://bioinfo.weizmann.ac.il/cards/
Genes and Disease (NCBI)	Organized by chromosome, provides descriptions of 60 diseases	▸http://www.ncbi.nlm.nih.gov/disease/
GeneClinics	A clinical information resource from the University of Washington, Seattle	▸http://www.geneclinics.org/
Genetic Alliance	International coalition of individuals, professionals, and genetic support organizations	▸http://www.geneticalliance.org/; search form ▸http://www.geneticalliance.org/diseaseinfo/search.html
Inherited Disease Genes Identified by Positional Cloning	From the National Human Genome Research Institute (NHGRI) at the NIH	▸http://genome.nhgri.nih.gov/clone/
The National Information Center for Children and Youth with Disabilities	An information and referral center in the United States	▸http://www.nichcy.org/
National Organization for Rare Disorders (NORD)	Federation of voluntary health organizations dedicated to helping people with rare "orphan" diseases and assisting the organizations that serve them	▸http://www.rarediseases.org/
Online Mendelian Inheritance in Man (OMIM)	Over 12,000 entries	▸http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM

You can see a list of positionally cloned genes at ▸http://genome.nhgri.nih.gov/clone/.

Jimenez-Sanchez et al. (2001), who analyzed 923 human genes that are associated with human disease. These genes primarily cause monogenic disorders, as expected since at present we know of relatively few genes that are mutated in complex disorders. They classified each disease gene according to the function of its protein product (Fig. 18.19*a*) Enzymes represent the largest functional category and account for 31% of the total gene products associated with disease. In contrast, only 15% of positionally cloned disease genes encode enzymes. Thus there may be some historical bias toward our knowledge of disease-causing mutations that are based on enzymatic defects.

Jimenez-Sanchez et al. (2001) further analyzed the correlation between the function of a gene product and the age of disease onset (Figs. 18.19*b–f*). Genes encoding enzymes and transcription factors are especially likely to be involved in disease in utero, reflecting the importance of transcription factors in early development. Enzymes are particularly involved in disease up to puberty (Figs. 18.9*b–d*). The developing fetus has access to its mother's metabolic systems and thus may be viable even if it has a gene defect. After birth, such diseases are manifested. Disease genes encoding enzymes are less prevalent in diseases having a later onset in life (Fig. 18.19*e*).

All of the common diseases in this sample occur with only a rare frequency when analyzed for any of four functional categories of disease—frequency, mode of inheritance, age of onset, and reduction of life expectancy (Fig. 18.20, leftmost column). This rare frequency reflects the population of disease genes that are currently

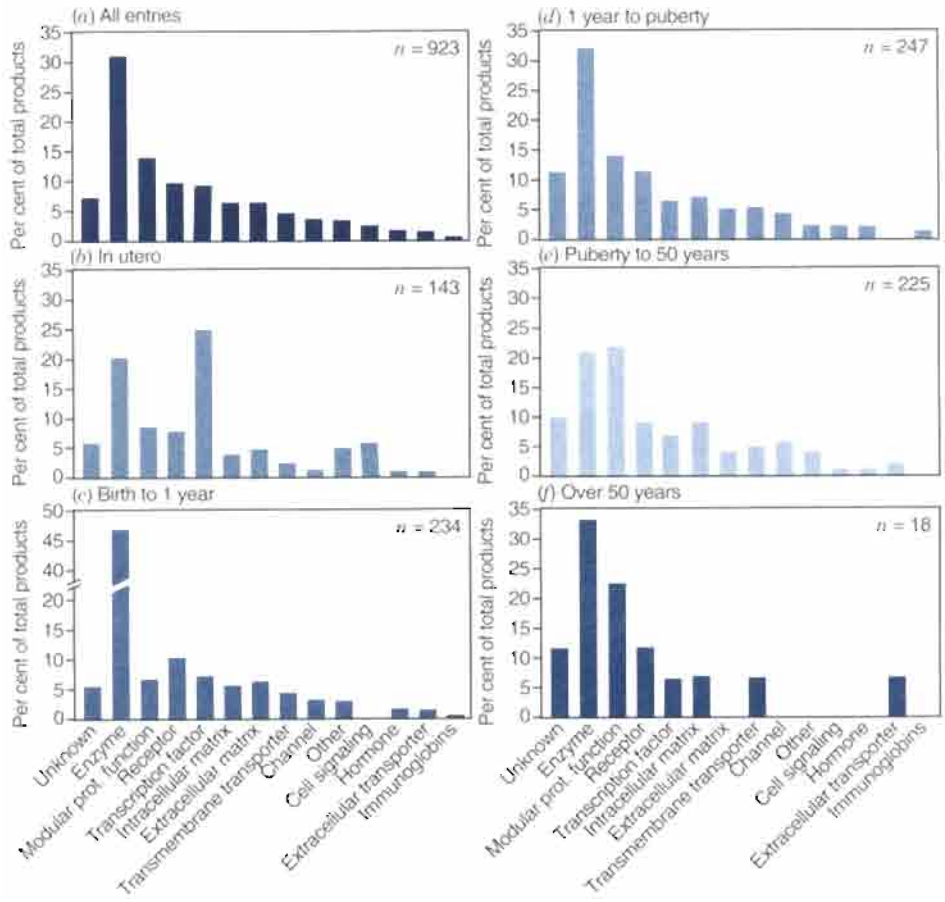


FIGURE 18.19. *The functions of the protein products of disease* genes *(from Jimenez-Sanchez et al., 2001): (a) all* genes (n = 923); *(b–f) disease* genes *listed according to the typical age of onset of the disease phenotype. Used with permission.*

available to study, that is, genes implicated in single-gene disorders. The mode of inheritance tends to be autosomal recessive, particularly for genes encoding enzymes. As described in Figure 18.19 as well, the age of onset tends to be in utero for transcription factors, from birth to one year for genes encoding enzymes, between one year and puberty and into adulthood for receptors, and early adulthood for modifiers of protein function (such as proteins that stabilize, activate, or fold other proteins). The severity of the disease, reflected in reduction of life expectancy, varies for diseases without a strong pattern based on functional categories.

These studies represent an early attempt to define a logic of disease. Such genomic-scale efforts will be enhanced when we have more information available on the genetic basis of complex disorders. Functional analyses may be combined using all the tools of bioinformatics and genomics to help elucidate the relationship between genotype and disease phenotype.

PERSPECTIVE

There are several kinds of bioinformatics approaches to human disease:

- Human disease is a consequence of variation in DNA sequence. These variations are catalogued in databases of molecular sequences (such as GenBank).

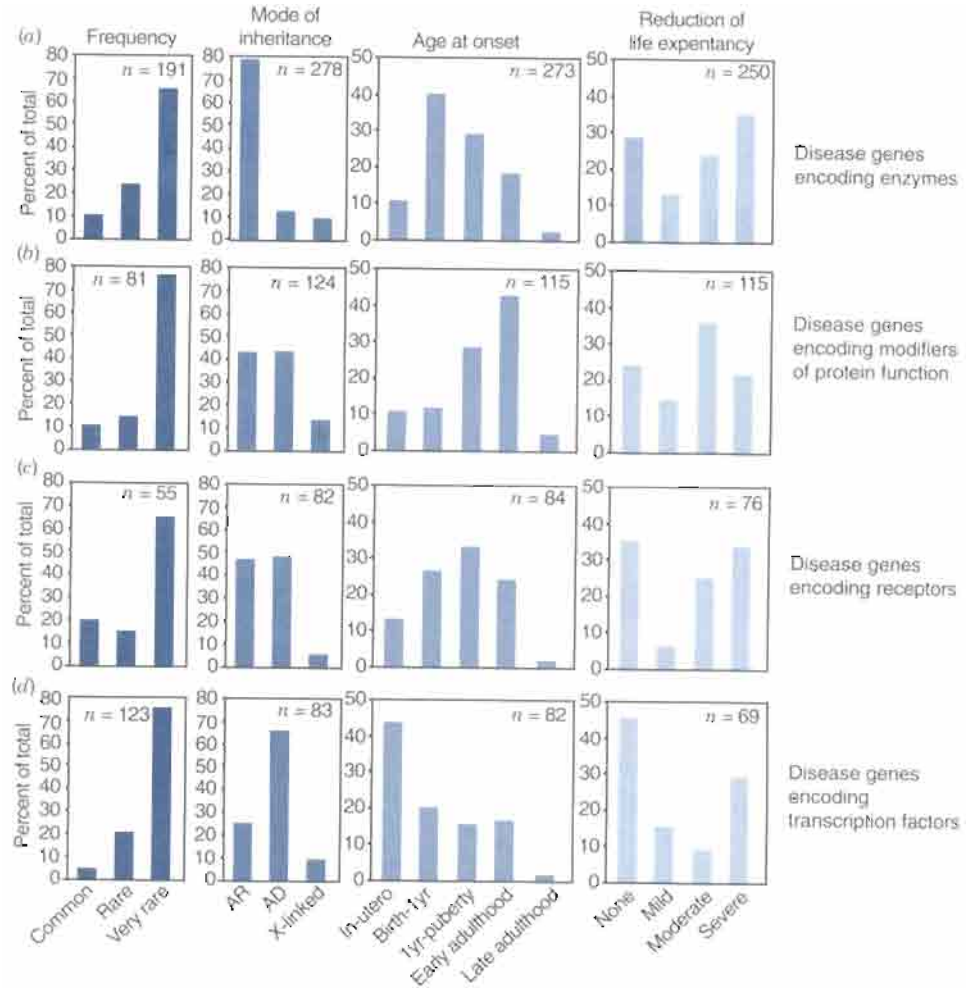


FIGURE 18.20. *The characteristics of diseases, organized by the function of the protein encoded by the disease gene. Abbreviations: AR, autosomal recessive; AD, autosomal dominant; early adulthood, puberty to <50 years old; late adulthood, >50 years old. From Jimenez-Sanchez et al. (2001). Used with permission.*

- Human disease databases have a major role in organizing information about disease genes. There are centralized databases, most notably OMIM, as well as locus-specific mutation databases.
- Functional genomics screens provide insight into the mechanisms of disease genes and disease processes.

PITFALLS

A fundamental gap in our understanding is how a genotype such as a mutated gene is related to a disease phenotype. We can approach disease from either end of the spectrum. Starting with a disease phenotype, we can ask what genes, when mutated, might cause this disease? Starting with a gene, we can ask what disease occurs when this gene is mutated? However, connecting these two ends of the continuum has been nearly impossible. For the majority of diseases, the discovery of a disease gene has not yet led to the subsequent discovery of new treatment options or cures, or to

an understanding of pathophysiology. Examples are muscular dystrophy and Rett syndrome. A hope is that bioinformatics and functional genomics approaches may lead to an understanding of biochemical pathways that account for the molecular basis of pathophysiology. This could be accomplished by learning the function of disease-causing genes in model organisms or through high-throughput technologies such as microarrays that describe the transcriptional response of susceptible cell types to the presence of a mutated gene.

WEB RESOURCES

TABLE 18-22 Web Resources for Study of Cancer

Resource	Description	URL
ACOR	Association of Cancer Online Resources	▸ http://www.acor.org/
Atlas of Genetics and Cytogenetics in Oncology and Haematology	A peer-reviewed on-line journal and cancer database	▸ http://www.infobiogen.fr/services/chromcancer/
Cancer Chromosome Aberration Project	Tools to define and characterize chromosomal alterations in cancer	▸ http://cgap.nci.nih.gov/Chromosomes/CCAP
The Cancer Gene Anatomy Project (CGAP)	At NCBI	▸ http://www.ncbi.nlm.nih.gov/ncicgap/
The Cancer Genome Project	The Wellcome Trust Sanger Institute	▸ http://www.sanger.ac.uk/CGP/
CancerNet	At the National Cancer Institute (of the NIH)	▸ http://cancernet.nci.nih.gov/
CancerWEB	Cancer resource site	▸ http://cancerweb.ncl.ac.uk/
Children's Cancer Web	Directory of childhood cancer resources	▸ http://www.CancerIndex.org/ccw/
Mitelman Database of Chromosome Aberrations in Cancer	Relates chromosomal aberrations to tumor characteristics	▸ http://cgap.nci.nih.gov/Chromosomes/Mitelman
OncoLink	Cancer information for the general public	▸ http://oncolink.upenn.edu/

DISCUSSION QUESTIONS

[18-1] Many neurological diseases such as Rett syndrome, vanishing white matter syndrome, and Huntington's disease have devastating consequences on brain function. For some of these diseases, the responsible genes have homologs in single-celled organisms such as fungi. Why do you think this is so?

[18-2] How have microarrays been used to study human disease? What are some specific examples of progress that has been made?

PROBLEMS

[18-1] *WT1* is a tumor suppressor gene lost in Wilms tumors and is also important for kidney development. Are there any SNPs within this gene and do they alter the protein sequence? What is the mutation distribution within the protein-coding region of *WT1*?

(a) Go to LocusLink and search for *WT1*. Click on the purple bar labeled VAR, which will take you to the page with SNPs in the *WT1* gene.
(b) Examine the SNPs. Do any alter the protein sequence?
(c) Try "Blast SNP" to find SNPs for this gene. Go to the site ▸ http://www.ncbi.nlm.nih.gov/SNP/ and

then click Blast SNP from the left sidebar. You can restrict the search to chromosome 11 (for *WT1*).

(d) Explore the *WT1* entry in OMIM.
____ What chromosome is the gene localized to?
____ Approximately how many allelic variants are known?
____ If you want to study Wilm's tumor, can you obtain cell lines from patients with this disorder? (*Hint:* Click "Coriell.")

(e) Look for known mutations in *WT1*. Go to the OMIM page on *WT1* and click on HGMD (Human Gene Mutation Database). This website places all mutations in various categories (missense, nonsense, deletion, etc.).
____ How many distinct phenotypes are caused by different mutations in this gene?
____ Based on the graphical mutation map (see link at the bottom of the HGMD page), do mutations in this gene tend to cluster or are they spread out evenly along the gene?

(f) What three disease genes are upstream of *WT1* on chromosome 11? What three disease genes are downstream of it? Is *WT1* on the short arm or the long arm of chromosome 11? Use OMIM to solve this problem.

(g) Find the three upstream and three downstream genes from *WT1* by using the human chromosome map site of NCBI. Note that this map shows all known genes (rather than just disease genes). You can get to the NCBI map viewer two ways:
- From the OMIM map entry for *WT1*, click "location."
- From the main page of NCBI, choose "human map viewer" from the right sidebar.

[18-2] How many inherited diseases have a known sequence associated with them?

(a) Do a LocusLink search for "disease_known AND has_seq." Make sure to limit the search to human (via the organism pop-up menu). This will retrieve the human LocusLink entries associated with a known disorder and sequence data.

(b) Go to OMIM and look up the statistics (on the left sidebar) to answer this question.

SELF-TEST QUIZ

[18-1] In humans, disorders that are inherited by simple Mendelian inheritance account for about what percentage of all human disease?
(a) 1%
(b) 10%
(c) 50%
(d) It is impossible to accurately measure the percentage.

[18-2] To a significant extent, susceptibility to an environmentally caused disorder such as poisoning from lead paint is determined by an individual's genes.
(a) true
(b) false

[18-3] Which of the following best describes single gene disorders? Each single gene disorder
(a) Is caused by a mutation in a single gene. They represent a basic category of disease that is in contrast to complex disorders.
(b) Is primarily caused by a mutation in a single gene, but the disease process always involve the contribution of many genes. Thus they represent a category of disease along a continuum with complex disorders.
(c) Is primarily caused by a mutation in a single gene in which the mutation almost always introduces a synonymous substitution.
(d) Is primarily caused by a mutation in a single gene in which the mutation almost always introduces a nonsynonymous substitution.

[18-4] In total, rare diseases in the United States affect about how many people?
(a) 200,000
(b) 2 million
(c) 25 million
(d) 100 million

[18-5] Single gene disorders tend to be
(a) rare in the general population, with an early onset in life
(b) common in the general population, with an early onset in life
(c) rare in the general population, with a late onset in life
(d) common in the general population, with a late onset in life

[18-6] There are several hundred locus-specific databases. What information do they offer that is not available in central databases such as OMIM and GeneCards?
(a) Comprehensive descriptions of the gene implicated in a disease
(b) Comprehensive lists of mutations associated with disease
(c) Links to foundations and other organizations
(d) Links to chromosome maps displaying the disease-causing gene

[18-7] Chromosomal disorders include monosomies, trisomies, and genomic rearrangements. Genomic microarrays can be used to detect such changes. These arrays consist of
(a) oligonucleotides that span genomic loci at about 1 to 10 megabase intervals
(b) cDNA clones that are assigned to known chromosomal loci
(c) BAC clones that are assigned to known chromosomal loci
(d) FISH probes

[18-8] Human disease genes have orthologs in a variety of organisms including worms, insects, and fungi. For a number

of human proteins that are implicated in disease, multiple sequence alignments with orthologous proteins have been made. These show that amino acid positions associated with disease-causing mutations in human proteins tend to be residues that are

(a) strongly conserved in other organisms
(b) sometimes conserved in other organisms
(c) poorly conserved in other organisms
(d) only sometimes aligned with orthologous sequences

Suggested Reading

An essential resource for the study of human disease is *The Metabolic and Molecular Basis of Inherited Disease* (Scriver et al. 2001). This four-volume tome has hundreds of chapters including introductions to disease from a variety of perspectives (e.g., Mendelian disorders, complex disorders, a logic of disease, mutation mechanisms, and animal models). A recommended introduction to disease is an essay by Barton Childs and David Valle (2000) in the inaugural volume of *Annual Review of Genomics and Human Genetics*.

The February 2001 issues of *Science* and *Nature* included brief articles on human disease by Leena Peltonen and Victor McKusick (2001) and Jimenez-Sanchez et al. (2001).

References

Ahringer, J. Turn to the worm! *Curr. Opin. Genet. Dev.* **7,** 410–415 (1997).

Allaman-Pillet, N., Djemai, A., Bonny, C., and Schorderet, D. F. Methylation status of CpG sites and methyl-CpG binding proteins are involved in the promoter regulation of the mouse Xist gene. *Gene Expr.* **7,** 61–73 (1998).

Amir, R. E., et al. Rett syndrome is caused by mutations in X-linked *MECP2,* encoding methyl-CpG-binding protein 2. *Nat. Genet.* **23,** 185–188 (1999).

Amir, R. E., and Zoghbi, H. Y. Rett syndrome: Methyl-CpG-binding protein 2 mutations and phenotype–genotype correlations. *Am. J. Med. Genet.* **97,** 147–152 (2000).

Anderson, R. N. Deaths: Leading causes for 1999. *Natl. Vital. Stat. Rep.* **49,** 1–87 (2001).

Antonarakis, S. E. Recommendations for a nomenclature system for human gene mutations. Nomenclature Working Group. *Hum. Mutat.* **11,** 1–3 (1998).

The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature* **408,** 796–815 (2000).

Armstrong, D. D. Review of Rett syndrome. *J. Neuropathol. Exp. Neurol.* **56,** 843–849 (1997).

Bailey, A., et al. Autism as a strongly genetic disorder: Evidence from a British twin study. *Psychol. Med.* **25,** 63–77 (1995).

Bailey, A., Phillips, W., and Rutter, M. Autism: Towards an integration of clinical, genetic, neuropsychological, and neurobiological perspectives. *J. Child Psychol. Psychiatry* **37,** 89–126 (1996).

Baker, P., Piven, J., Schwartz, S., and Patil, S. Brief report: Duplication of chromosome 15q11-13 in two individuals with autistic disorder. *J. Autism Dev. Disord.* **24,** 529–535 (1994).

Barton, M., Volkmar, F. How commonly are known medical conditions associated with autism? *J. Autism Dev. Disord.* **28,** 273–278, (1998).

Bauman, M. L., Kemper, T. L., and Arin, D. M. Microscopic observations of the brain in Rett syndrome. *Neuropediatrics* **26,** 105–108 (1995).

Beaudet, A. L., Scriver, C. R., Sly, W. S., and Valle, D. Genetics, biochemistry, and molecular bases of variant human phenotypes. In Scriver et al. (Eds.), *The Metabolic & Molecular Bases of Inherited Disease,* Vol. 1. Mc Graw-Hill, New York, 2001, pp. 3–45.

Belichenko, P. V., Hagberg, B., and Dahlstrom, A. Morphological study of neocortical areas in Rett syndrome. *Acta Neuropathol. (Berl.)* **93,** 50–61 (1997).

Beroud, C., Collod-Beroud, G., Boileau, C., Soussi, T., and Junien, C. UMD (Universal mutation database): a generic software to build and analyze locus-specific databases. *Hum. Mutat.* **15,** 86–94, 2000.

Bienvenu, T., et al. *MECP2* mutations account for most cases of typical forms of Rett syndrome. *Hum. Mol. Genet.* **9,** 1377–1384 (2000).

Bono, H., Kasukawa, T., Furuno, M., Hayashizaki, Y., and Okazaki, Y. FANTOM DB: Database of Functional Annotation of RIKEN Mouse cDNA Clones. *Nucleic Acids Res.* **30,** 116–118 (2002).

Brown, A. F., and McKie, M. A. MuStaR and other software for locus-specific mutation databases. *Hum. Mutat.* **15,** 76–85 (2000).

Buetow, K. H., Edmonson, M. N., and Cassidy, A. B. Reliable identification of large numbers of candidate SNPs from public EST data. *Nat. Genet.* **21,** 323–325 (1999).

Bundey, S., Hardy, C., Vickers, S., Kilparrick, M. W., and Corbett, J. A. Duplication of the 15q11-13 region in a patient with

autism, epilepsy and ataxia. *Dev. Med. Child. Neurol.* **36,** 736–742 (1994).

Cedar, H. DNA methylation and gene activity. *Cell* **53,** 3–4 (1988).

Cheadle, J. P., et al. Long-read sequence analysis of the *MECP2* gene in Rett syndrome patients: Correlation of disease severity with mutation type and location. *Hum. Mol. Genet.* **9,** 1119–1129 (2000).

Childs, B., and Valle, D. *Genetics, Biology and Disease.* Annual Reviews, Palo Alto, CA, 2000, pp. 1–19.

Ciaranello, A. L., and Ciaranello, R. D. The neurobiology of infantile autism. *Annu. Rev. Neurosci.* **18,** 101–128 (1995).

Claustres, M., Horaitis, O., Vanevski, M., and Cotton, R. G. Time for a unified system of mutation description and reporting: A review of locus-specific mutation databases. *Genome Res.* **12,** 680–688 (2002).

Colantuoni, C., Purcell, A. E., Bouton, C. M., and Pevsner, J. High throughput analysis of gene expression in the human brain. *J. Neurosci. Res.* **59,** 1–10 (2000).

Collins, F. S., Brooks, L. D., and Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8,** 1229–1231, (1998).

Cook, E. H., Jr. et al. Autism or atypical autism in maternally but not paternally derived proximal 15q duplication. *Am. J. Hum. Genet.* **60,** 928–934 (1997).

Cook, E. H., Jr. et al. Linkage-disequilibrium mapping of autistic disorder, with 15q11-13 markers. *Am. J. Hum. Genet.* **62,** 1077–1083 (1998).

Cowles, C. R., Joel, N. H., Altshuler, D., and Lander, E. S. Detection of regulatory variation in mouse genes. *Nat. Genet.* **32,** 432–437 (2002).

den Dunnen, J. T., and Antonarakis, S. E. Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Hum. Mutat.* **15,** 7–12 (2000).

DiMauro, S., and Schon, E. A. Mitochondrial DNA mutations in human disease. *Am. J. Med. Genet.* **106,** 18–26 (2001).

Dipple, K. M., and McCabe, E. R. Modifier genes convert "simple" Mendelian disorders to complex traits. *Mol. Genet. Metab.* **71,** 43–50, (2000).

Dipple, K. M., Phelan, J. K., and McCabe, E. R. Consequences of complexity within biological networks: Robustness and health, or vulnerability and disease. *Mol. Genet. Metab.* **74,** 45–50 (2001).

Flejter, W. L., et al. Cytogenetic and molecular analysis of inv dup(15) chromosomes observed in two patients with autistic disorder and mental retardation. *Am. J. Med. Genet.* **61,** 182–187 (1996).

Fombonne, E. The epidemiology of autism: A review. *Psychol. Med.* **29,** 769–786 (1999).

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229 (2002).

Garrod, A. E. *Inborn errors of metabolism: The Croonian Lectures delivered before the Royal College of Physicians of London, in June, 1908.* Frowde, Hodder and Stoughton, London (1909).

Garrod, A. E. *Inborn factors in disease: An essay.* Clarendon Press, Oxford (1931).

Garrod, A. E. The incidence of alkaptonuria: A study in chemical individuality. *Lancet* **ii,** 1616–1620 (1902).

Geerdink, N., et al. *MECP2* mutation in a boy with severe neonatal encephalopathy: Clinical, neuropathological and molecular findings. *Neuropediatrics* **33,** 33–36 (2002).

Gillberg, C., and Wing, L. Autism: Not an extremely rare disorder. *Acta Psychiatr. Scand.* **99,** 399–406 (1999).

Goodstadt, L., and Ponting, C. P. Sequence variation and disease in the wake of the draft human genome. *Hum. Mol. Genet.* **10,** 2209–2214 (2001).

Guttmacher, A. E. Human genetics on the web. *Annu. Rev. Genomics Hum. Genet.* **2,** 213–233 (2001).

Hagberg, B., Aicardi, J., Dias, K., and Ramos, O. A progressive syndrome of autism, dementia, ataxia, and loss of purposeful hand use in girls: Rett's syndrome: report of 35 cases. *Ann. Neurol.* **14,** 471–479 (1983).

Hammer, S., Dorrani, N., Dragich, J., Kudo, S., and Schanen, C. The phenotypic consequences of *MECP2* mutations extend beyond Rett syndrome. *Ment. Retard. Dev. Disabil. Res. Rev.* **8,** 94–98 (2002).

Hamosh, A., Scott, A. F., Amberger, J., Valle, D., and McKusick, V. A. Online Mendelian Inheritance in Man (OMIM). *Hum. Mutat.* **15,** 57–61 (2000).

Hamosh, A., et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **30,** 52–55 (2002).

Hermsen, M. A., Weiss, M. M., Meijer, G. A., and Baak, J. P. Detection of chromosomal abnormalities by comparative genomic hybridization. *Methods Mol. Biol.* **175,** 47–55 (2001).

Hill, A. V., The genomics and genetics of human infectious disease susceptibility. *Annu. Rev. Genomics Hum. Genet.* **2,** 373–400 (2001).

Holt, I. J., Harding, A. E., and Morgan-Hughes, J. A. Deletions of muscle mitochondrial DNA in patients with mitochondrial myopathies. *Nature* **331,** 717–719, 1988.

Hrabe de Angelis, M., and Strivens, M. Large-scale production of mouse phenotypes: The search for animal models for inherited diseases in humans. *Brief. Bioinform.* **2,** 170–180 (2001).

Huppke, P., Laccone, F., Kramer, N., Engel, W., and Hanefeld, F. Rett syndrome: Analysis of *MECP2* and clinical characterization of 31 patients. *Hum. Mol. Genet.* **9,** 1369–1375 (2000).

International Molecular Genetic Study of Autism Consortium. A full genome screen for autism with evidence for linkage

to a region on chromosome 7q. *Hum. Mol. Genet.* **7,** 571–578 (1998).

Jellinger, K., Armstrong, D., Zoghbi, H. Y., and Percy, A. K. Neuropathology of Rett syndrome. *Acta Neuropathol.* **76,** 142–158 (1988).

Jimenez-Sanchez, G., Childs, B., and Valle, D. Human disease genes. *Nature* **409,** 853–855 (2001).

Kiberstis, P., and Roberts, L. It's not just the genes. *Science* **296,** 685 (2002).

Kim, S. J., and Cook, E. H., Jr. Novel de novo nonsense mutation of *MECP2* in a patient with Rett syndrome. *Hum. Mutat.* **15,** 382–383 (2000).

Knoppers, B. M., and Laberge, C. M. Ethical guideposts for allelic variation databases. *Hum. Mutat.* **15,** 30–35 (2000).

Krawczak, M., and Cooper, D. N. The human gene mutation database. *Trends Genet.* **13,** 121–122 (1997).

Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22,** 139–144 (1999).

Kruglyak, L., and Nickerson, D. A. Variation is the spice of life. *Nat Genet.* **27,** 234–236 (2001).

Lamb, J. A., Moore, J., Bailey, A., and Monaco, A. P. Autism: Recent molecular genetic advances. *Hum. Mol. Genet.* **9,** 861–868 (2000).

Lichter, P., Joos, S., Bentz, M., and Lampel, S. Comparative genomic hybridization: Uses and limitations. *Semin. Hematol.* **37,** 348–357 (2000).

Lupski, J. R. Genomic disorders: Structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* **14,** 417–422 (1998).

Miller, M. P., and Kumar, S. Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.* **10,** 2319–2328 (2001).

Murray, C. J. L., and Lopez, A. D. (eds.) *The Global Burden of Disease.* Harvard University Press, Cambridge (1996).

Naidu, S. Rett syndrome: A disorder affecting early brain growth. *Ann. Neurol.* **42,** 3–10 (1997).

Nan, X., Campoy, F. J., and Bird, A. MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. *Cell* **88,** 471–481 (1997).

Nass, S., and Nass, M. M. K. Intramitochondrial fibers with DNA characteristics. *J. Cell Biol.* **19,** 613–629 (1963).

Ng, H. H., and Bird, A. DNA methylation and chromatin modification. *Curr. Opin. Genet. Dev.* **9,** 158–163 (1999).

Nolan, P. M., Hugill, A., and Cox, R. D. ENU mutagenesis in the mouse: Application to human genetic disease. *Briefings Functional Genomics Proteomics* **1,** 278–289 (2002).

Olsson, I., Steffenburg, S., and Gillberg, C. Epilepsy in autism and autisticlike conditions. A population-based study. *Arch. Neurol.* **45,** 666–668 (1988).

Patil, N., Berno, A. J., Hinds, D. A., Barrett, W. A., Doshi, J. M., Hacker, C. R., Kautzer, C. R., Lee, D. H., Marjoribanks, C., McDonough, D. P., Nguyen, B. T., Norris, M. C., Sheehan, J. B., Shen, N., Stem, D., Stokowski, R. P., Thomas, D. J., Trulson, M. O., Vyas, K. R., Frazer, K. A., Fodor, S. P., and Cox, D. R. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294,** 1719–1723 (2001).

Pauling, L., Itano, H. A., Singer, S. J., and Wells, I. C. Sickle cell anemia, a molecular disease. *Science* **110,** 543–548 (1949).

Peltonen, L., and McKusick, V. A. Genomics and medicine. Dissecting human disease in the postgenomic era. *Science* **291,** 1224–1229 (2001).

Philippe, A., et al. Genome-wide scan for autism susceptibility genes. Paris Autism Research International Sibpair Study. *Hum. Mol. Genet.* **8,** 805–812 (1999).

Phillips, M. S., Lawrence, R., Sachidanandam, R., Morris, A. P., Balding, D. J., Donaldson, M. A., Studebaker, J. F., Ankener, W. M., Alfisi, S. V., Kuo, F. S., Camisa, A. L., Pazorov, V., Scott, K. E., Carey, B. J., Faith, J., Katari, G., Bhatti, H. A., Cyr, J. M., Derohannessian, V., Elosua, C., Forman, A. M., Grecco, N. M., Hock, C. R., Kuebler, J. M., Lathrop, J. A., Mockler, M. A., Nachtman, E. P., Restine, S. L., Varde, S. A., Hozza, M. J., Gelfand, C. A., Broxholme, J., Abecasis, G. R., Boyce-Jacino, M. T., and Cardon, L. R. Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat Genet.* **33,** 382–387 (2003).

Pinkel, D., et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20,** 207–211 (1998).

Piven, J. The biological basis of autism. *Curr. Opin. Neurobiol.* **7,** 708–712 (1997).

Pollack, J. R., et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23,** 41–46 (1999).

Purcell, A. E., Jeon, O. H., Zimmerman, A. W., Blue, M. E., and Pevsner, J. Postmortem brain abnormalities of the glutamate neurotransmitter system in autism. *Neurology* **57,** 1618–1628 (2001).

Rapin, I. Autism. *N. Engl. J. Med.* **337,** 97–104 (1997).

Rapin, I., and Katzman, R. Neurobiology of autism. *Ann. Neurol.* **43,** 7–14 (1998).

Razin, A. CpG methylation, chromatin structure and gene silencing–a three-way connection. *EMBO J* **17,** 4905–4908 (1998).

Reiter, L. T., Potocki, L., Chien, S., Gribskov, M., and Bier, E. A systematic analysis of human disease-associated gene sequences in *Drosophila melanogaster. Genome Res.* **11,** 1114–1125 (2001).

Rossi, P. G., Parmeggiani, A., Bach, V., Santucci, M., and Visconti, P. EEG features and epilepsy in patients with autism. *Brain Dev.* **17,** 169–174 (1995).

Rubin, G. M., et al. Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).

Safran, M., et al. GeneCards(TM) 2002: Towards a complete, object-oriented, human gene compendium. *Bioinformatics* **18,** 1542–1543 (2002).

Schon, E. A. Mitochondrial genetics and disease. *Trends Biochem. Sci.* **25,** 555–560 (2000).

Scriver, C., Beaudet, A., Sly, W., and Valle, D. (Eds.). *The Metabolic and Molecular Basis of Inherited Disease.* McGraw-Hill, New York (2001).

Scriver, C. R., and Childs, B. *Garrod's Inborn Factors in Disease.* New York, Oxford University Press (1989).

Scriver, C. R., Nowacki, P. M., and Lehvaslaiho, H. Guidelines and recommendations for content, structure, and deployment of mutation databases. *Hum. Mutat.* **13,** 344–350 (1999).

Scriver, C. R., Nowacki, P. M., and Lehvaslaiho, H. Guidelines and recommendations for content, structure, and deployment of mutation databases: II. Journey in progress. *Hum. Mutat.* **15,** 13–15 (2000).

Sherry, S. T., Ward, M., and Sirotkin, K. Use of molecular variation in the NCBI dbSNP database. *Hum. Mutat.* **15,** 68–75 (2000).

Singer, C. *The Fasciculo di Medicina Venice 1493.* R. Lier and Co., Florence, 1925.

Smalley, S. L., Asarnow, R. F., and Spence, M. A. Autism and genetics. A decade of research. *Arch. Gen. Psychiatry* **45,** 953–961 (1988).

Szatmari, P., Jones, M. B., Zwaigenbaum, L., and MacLean, J. E. Genetics of autism: Overview and new directions. *J. Autism Dev. Disord.* **28,** 351–368 (1998).

Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L., and Kwok, P. Y. Overlapping genomic sequences: A treasure trove of single-nucleotide polymorphisms. *Genome Res.* **8,** 748–754 (1998).

Thomas, C. L. (ed.) *Taber's Cyclopedic Medical Dictionary.* F. A. Davis Company, Philadelphia, 1997.

Todd, J. A. Multifactorial diseases: Ancient gene polymorphism at quantitative trait loci and a legacy of survival during our evolution. In C. Scrivner et al. (Eds.), *The Metabolic & Molecular Bases of Inherited Disease,* Vol. 1. Mc Graw-Hill, New York, 2001, pp. 193–201.

Turner, M., Barnby, G., and Bailey, A. Genetic clues to the biological basis of autism. *Mol. Med. Today* **6,** 238–244 (2000).

Van den Veyver, I. B., and Zoghbi, H. Y. Methyl-CpG-binding protein 2 mutations in Rett syndrome. *Curr. Opin. Genet. Dev.* **10,** 275–279 (2000).

Volkmar, F. Recently diagnosed with autism, autism or not. *J. Autism Dev. Disord.* **28,** 269–270 (1998).

Volkmar, F. R., and Nelson, D. S. Seizure disorders in autism. *J. Am. Acad. Child. Adolesc. Psychiatry* **29,** 127–129 (1990).

Wallace, D. C., Singh, G., Lott, M. T., Hodge, J. A., Schurr, T. G., Lezza, A. M., Elsas, L. J. 2nd, Nikoskelainen, E. K. Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science* **242,** 1427–1430, 1988a.

Wallace, D. C., Zheng, X. X., Lott, M. T., Shoffner, J. M., Hodge, J. A., Kelley, R. I., Epstein, C. M., Hopkins, L. C. Familial mitochondrial encephalomyopathy (MERRF): genetic, pathophysiological, and biochemical characterization of a mitochondrial DNA disease: *Cell* **55,** 601–610, 1988b.

Wan, M., et al. Rett syndrome and beyond: Recurrent spontaneous and familial *MECP2* mutations at CpG hotspots. *Am. J. Hum. Genet.* **65,** 1520–1529 (1999).

Wang, Z., and Moult, J. SNPs, protein structure, and disease. *Hum. Mutat.* **17,** 263–270 (2001).

Waterhouse, L. Genes tPA, Fyn, and FAK in autism? *J. Autism Dev. Disord.* **27,** 220–223 (1997).

Waterston, R. H., et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562 (2002).

Wood, V., et al. The genome sequence of *Schizosaccharomyces pombe. Nature* **415,** 871–880 (2002).

Xiang, F., et al. Mutation screening in Rett syndrome patients. *J. Med. Genet.* **37,** 250–255 (2000).

Zeev, B. B., et al. Rett syndrome: Clinical manifestations in males with *MECP2* mutations. *J. Child Neurol.* **17,** 20–24 (2002).

Zuckerlandl, E., and Pauling, L. Molecular disease, evolution, and genic heterogeneity. In M. Kasha, and B. Pullman, (Eds.), *Horizons in Biochemistry,* Albert Szent-Gyorgyi Dedicatory Volume. Academic Press, New York (1962).

Epilogue

Biology has entered the genomic era, in which the sequencing and analysis of genomes introduce a new approach to biology. The first third of this book focused on the acquisition of information, especially DNA and protein sequence data. We then compared sequences either directly to each other or to databases. In the middle third of this book, we considered gene expression and then proteins and proteomics, including an introduction to molecular evolution. Finally, in the last third we explored genomes across the tree of life, often with a perspective on human disease.

We are at a phase of biological science when the Human Genome Project is essentially complete, and hundreds of other genomes have been sequenced. A group of scientists at the National Human Genome Research Institute at the N.I.H.—including Francis Collins, Eric Green, Alan Guttmacher, and Mark Guyer—recently presented their views on the next phase of genomics (Collins et al., 2003). After two years of discussions with hundreds of scientists, they listed 15 challenges that constitute a framework for a vision for the future of genomics. These challenges connect three major themes: genomics and biology, genomics and health, and genomics and society. The list is as follows:

You can read more about these 15 challenges, and see a copy of the Collins et al. (2003) paper, at ▸http://www.genome.gov/About/Planning/. The list is adapted from Collins et al. (2003) and the U.S. National Human Genome Research Institute (used with permission).

I Genomics and Biology

- Comprehensively identify the structural and functional components encoded in the human genome
- Elucidate the organization of genetic networks and protein pathways and establish how they contribute to cellular and organismal phenotypes
- Develop a detailed understanding of the heritable variation in the human genome
- Understand evolutionary variation across species and the mechanisms underlying it
- Develop policy options that facilitate the widespread use of genome information in both research and clinical settings

II Genomics and Health

- Develop robust strategies for identifying the genetic contributions to disease and drug response
- Develop strategies to identify gene variants that contribute to good health and resistance to disease
- Develop genome-based approaches to prediction of disease susceptibility and drug response, early detection of illness, and molecular taxonomy of disease states

- Use new understanding of genes and pathways to develop powerful new therapeutic approaches to disease
- Investigate how genetic risk information is conveyed in clinical settings, how that information influences health strategies and behaviors, and how these affect health outcomes and costs
- Develop genome-based tools that improve the health of all

III Genomics and Society

- Develop policy options for the uses of genomics in medical and non-medical settings
- Understand the relationships between genomics, race and ethnicity, and the consequences of uncovering these relationships
- Understand the consequences of uncovering the genomic contributions to human traits and behaviors
- Assess how to define the ethical boundaries for uses of genomics

All these goals are extremely broad. To achieve any of them requires a realistic understanding of the tools of bioinformatics and genomics. It is hoped that the descriptions of molecular sequences, databases, algorithms, genomes, biological principles, and other topics presented in this book will contribute in some small way to the ability of students and researchers to work towards these broad goals.

REFERENCE

Collins, F. S., Green, E. D., Guttmacher, A. E., and Guyer, M. S. A vision for the future of genomics research. *Nature* **422,** 835–847, 2003.

Appendix
GCG for Protein and DNA Analysis

Topic 1: Introduction to GCG

The GCG (Genetics Computer Group) package is a collection of more than 130 interrelated software programs used to analyze protein, RNA, and DNA sequences. It is a commercial product that is similar in scope to Vector NTI (by Informax) and to the freely available, UNIX-based tools of Emboss (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/).

Information on GCG is available at http://www.gcg.com (or http://www.accelrys.com/about/gcg.html). GCG was founded in the Department of Genetics at the University of Wisconsin-Madison in 1982. It is used by more than 40,000 scientists at over 650 institutions around the world.

GCG topics include the following:

Comparison	Importing and exporting
Database searching	Mapping
DNA/RNA secondary structure	Primer selection
Editing and publication	Protein analysis
Evolution	SeqLab
Fragment assembly	Translation
Gene finding and pattern recognition	Utilities

Topic 2: General Commands

UNIX General Commands

When you use the GCG suite of programs, you are in a UNIX environment. Some useful commands follow:

> cat x	Shows the document x
> more x	Shows the document x one page at a time (press spacebar to advance)
> ls	Lists the files in a directory
> ls *.pep	Lists all the files in a directory that end with the suffix .pep
> ^C	Interrupts what you are doing and returns to the command line

> ^S	Pauses the output on the screen
> ^Q	Resumes the output
> man	On-line help manual

GCG General Commands

Within GCG, there are several important commands to know.

Getting Help

For any program you are working on, it is very useful to read through the help menu. Typically, this menu provides features such as the following:

- An overview of what the program does
- An example of how to actually use it
- A description of the algorithm used, together with literature references
- A comparison of what this program does to what related GCG programs can do

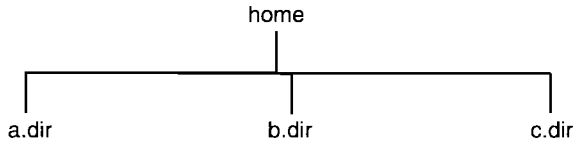
Some useful commands on getting help are

> genhelp	Gets help on all the 130 programs within GCG
> genhelp gap	Gets help on the program called GAP

Organizing Your Files into Directories

One can quickly accumulate dozens (or hundreds) of GCG files. A good way to organize your files is to make a separate directory for each project you are working on (Figure Appendix 1). When you want to make a directory, simply type "mkdir" followed by the name of the new directory. Then to get to that directory in order to work in it, type "down a.dir." You can move between directories laterally by typing "over any.dir." To get back to your main, starting directory, you can always type "home." If you want to confirm where you are at any time, you can always type "ls" to list whatever files you have and to see the name of your current directory or subdirectory. A summary of commands used to organize your files are

> mkdir	Creates a new directory
> mkdir a.dir	Creates a new directory called a.dir
> down a.dir	Moves from the home directory down to a.dir
> over b.dir	Moves over to the directory called b.dir
> up	Moves up to a higher directory.
> home	Goes to your main, home directory.



Naming Files

A handy rule is to name any protein file you work with the suffix ".pep" (for peptide; for example, actin.pep). Name any DNA file with the suffix ".seq" (for sequence; for example, actin.seq). You will typically get the chance to name a file when you first

create it in the main sequence editor, seqed (see below). In the GCG programs names such as a.pep or a.seq should be used to ensure that the appropriate data are being entered into a program (e.g., .seq files for restriction enzyme analyses or .pep files for hydropathy profiles). There are some exceptions; for example, a DNA sequence that has been flipped to show bottom-strand sequence should have the suffix . rev.

Deleting, Renaming, and Copying Files

When you are deleting files, be very careful! It is easy to delete, and file recovery is not easy. Type "> rm x" to remove a file called x (rm stands for remove):

> rm x	Deletes a file called x
> rm *z*	Deletes all the files with a z in the name
> rm *.pep	Deletes all your peptide/protein files with the suffix .pep

To rename a file, use the mv (move) command:

> mv somename.pep betternname.pep	Renames a file somename.pep as a file called betternname.pep

To copy a file, use the cp (copy) command:

> cp *.pep ~ /somedirectory/	Copies all files containing the suffix .pep to a directory called somedirectory

Note that the tilde ~ represents the home directory.

TOPIC 3: ENTERING AND EDITING SEQUENCES

The main feature of GCG is its ability to manipulate sequences. In most cases, you need to begin by importing some sequence into GCG. Typical examples of what you would want import include the following:

- You get some DNA sequence from a Core Facility on your computer, and you want to know how well it matches a known DNA sequence or a known protein sequence.
- You directly sequence a protein and want to analyze its properties.
- Someone sends you a recombinant fusion protein and you want to know its properties.
- You need to know if two proteins are significantly homologous or not.

In many cases, you will begin by finding a protein or DNA sequence through the National Center for Biotechnology Information (NCBI) or SwissProt (Chapter 2).

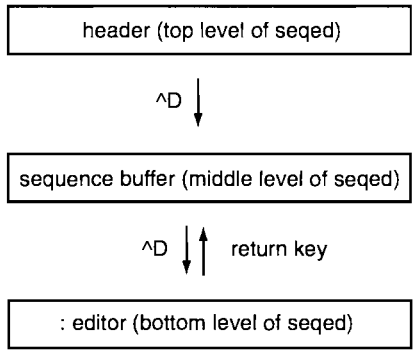
Copy your DNA or protein sequence, then paste it into GCG using the sequence editor seqed:

> seqed x.pep	Makes (or edits) a protein file called x.pep
> seqed y.seq	Makes (or edits) a DNA file called y.seq

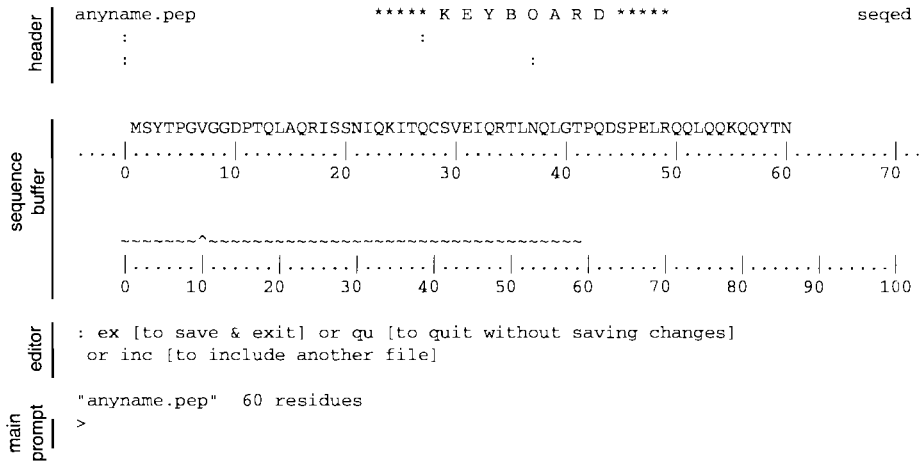
You now find yourself in the seqed program. On the page, there are three levels: a header at the top, a sequence buffer in the middle, and a command line editor down below. Type ^D (control D) to move down to the sequence buffer, then "paste" in your sequence.

Commonly used editor commands include the following:

: exit	Ends a seqed session and save your changes
: quit	Ends a seqed session without saving changes
: include	Pastes in portion of another DNA sequence
: help	Gets help on seqed



The figure below shows an example of these levels in the seqed program.



To enter human retinol-binding protein (RBP4) in GCG, type the following (hsrbp stands for *Homo sapiens* retinol-binding protein): >seqed hsrbp.pep <return>. Note that the GCG cursor is now in the top header region:

```
hsrbp.pep          ***** K E Y B O A R D *****          seqed
    :                                                       :
    :                                                       :
    :                                                       :
    :                                                       :
```

In a browser such as Netscape, open the GenBank protein record for RBP. Copy the accession number and any other relevant information into the header region. When you are finished, press ^D to enter the main sequence buffer. Copy and paste the amino acid sequence of RBP into GCG (note that GCG will ignore numbers that are associated with the letters of protein or nucleic acid sequences):

```
test.pep                ***** K E Y B O A R D *****                 seqed
    : LOCUS      NP_006735     199 aa                    PRI       08-JUN0 :
    : DEFINITION retinol-binding protein 4, interstitial precursor [Homo . :
    : ACCESSION  NP_006735
    : PID        g5803139

crllnldgtcadsysfvfsrdpnglppeaqkivrqrqeelclarqyrlivhngycdgrsernll
....|.........|.........|.........|.........|.........|.........|.........|....
  140       150       160       170       180       190       200       210

    ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~^
    |......|......|......|......|......|......|......|......|......|......|
    0      20     40     60     80     100    120    140    160    180    200
```

Next, in GCG press ^D to move down from the sequence buffer to the command line. The command line has a colon (":"). Type "ex" (for "exit"):

```
: ex

You have successfully created the file hsrbp.pep.
Type > ls to confirm that this file has appeared in your directory.
Type > cat hsrbp.pep to see this file displayed:

> cat hsrbp.pep
!!AA_SEQUENCE 1.0
LOCUS       NP_006735 199 AA PRI 08-JUN-2000
DEFINITION  RETINOL-BINDING PROTEIN 4, INTERSTITIAL PRECURSOR
             [HOMO SAPIENS].
ACCESSION   NP_006735
PID         G5803139
VERSION     NP_006735.1 GI:5803139
DBSOURCE    REFSEQ: ACCESSION NM_006744.2
KEYWORDS    .
SOURCE      HUMAN.
ORGANISM    HOMO SAPIENS

HSRBP.PEP LENGTH: 199  AUGUST 30, 2000 19:19  TYPE: P  CHECK: 8315 ..

       1 MKWVWALLLL AAWAAAERDC RVSSFRVKEN FDKARFSGTW YAMAKKDPEG

      51 LFLQDNIVAE FSVDETGQMS ATAKGRVRLL NNWDVCADMV GTFTDTEDPA

     101 KFKMKYWGVA SFLQKGNDDH WIVDTDYDTY AVQYSCRLLN LDGTCADSYS

     151 FVFSRDPNGL PPEAQKIVRQ RQEELCLARQ YRLIVHNGYC DGRSERNLL
>
```

Optionally, you can use the following commands:

> onecase	make sequences all upper or lowercase
> onecase *.pep	make all protein sequences one case

In this appendix, we will also use the example of syntaxins. These are proteins that bind vesicle proteins and coordinate the docking of vesicles with the appropriate target membrane.

Topic 4: Pairwise Alignment

BestFit	Compares two proteins or DNA sequences; focus on a local alignment
GAP	Compares two proteins or DNA sequences; focus on a global alignment
TFastA	Compares one protein to the six proteins potentially encoded from one DNA sequence

Example of How to Use BestFit Program

In this example, compare two proteins (human syntaxin 1a and human syntaxin 7) by typing "bestfit," then the names of the proteins to be compared (with a space between them):

```
> bestfit humsyn1a.pep humsyn7.pep

BestFit makes an optimal alignment of the best segment of similarity
between two sequences. Optimal alignments are found by inserting gaps
to maximize the number of matches using the local homology algorithm
of Smith and Waterman.

                Begin (* 1 *) ?
              End (*   289 *) ?

                Begin (* 1 *) ?
              End (*   261 *) ?

What is the gap creation penalty (* 12 *) ?

What is the gap extension penalty (* 4 *) ?

What should I call the paired output display file (* humsyn1a.pair *) ?

Aligning .............-.
Aligning .............-..

         Gaps:      2
      Quality:    194
Quality Ratio:  0.911
 % Similarity: 39.151
       Length:    216
```

Next, use the cat command to view the alignment.

```
> cat humsyn1a.pair
 BESTFIT of:  humsyn1a.pep  check:  5551  from:  1  to:  289

LOCUS       HUMSYN1A     2088 bp    mRNA          PRI 23-NOV-1994
DEFINITION  Human syntaxin 1A mRNA, complete cds.
ACCESSION   L37792
NID         g577487
KEYWORDS    syntaxin 1A.
SOURCE      Homo sapiens male adult brain cDNA to mRNA . . .

 to: humsyn7.pep  check: 338  from: 1  to: 261

 Symbol comparison table: /prog/GCG/gcgcore/data/rundata/blosum62.cmp
 CompCheck: 6430

       Gap Weight:    12          Average Match:  2.912
    Length Weight:     4       Average Mismatch: -2.003
          Quality:   194                 Length:    216
            Ratio: 0.911                   Gaps:      2
Percent Similarity: 39.151      Percent Identity: 26.887

      Match display thresholds for the alignment(s):
                  | =  IDENTITY
                  : =    2
                  . =    1

humsyn1a.pep x humsyn7.pep October 28, 1998 19:27  ..
```

```
          .          .          .          .          .
 71 TKEELEELMSDIKKTANKVRSKLKSIEQSIEQEEGLNRSSADLRIRKTQH 120
    | ::  ||   :..          |  :. |.:   |  . .: | || |
 41 TPQDSPELRQQLQQKQQYTNQLAKETDKYIKEFGSLPTTPSEQRQRKIQK 90
          .          .          .          .          .
121 STLSRKFVEVMSEYNATQSDYRERCK...GRIQRQLEITGRTTTSEELED 167
      |  .    :. :   |    || |    |:.    :.|        |
 91 DRLVAECTTSLTNFQKVQRQAAEREKEFVARVRASSRVSGSFPEDSSKER 140
          .          .          .          .          .
168 MLESGNPAIFASGIIMDSSISKQALSEIETRHSEIIKLENSIRELHDMFM 217
     | |          : |  |..  |  |  | | | .||  | ::.:.|
141 NLVSWESQTQPQVQVQDEEITEDDLRLIHERESSIRQLEADIMDINEIFK 190
          .          .          .          .          .
218 DMAMLVESQGEMIDRIEYNVEHAVDYVERAVSDTKKAVKYQSKARRKKIM 267
    |: |::  ||:.|| || |||.|  :|:.|     :|  || |. || :
191 DLGMMIHEQGDVIDSIEANVENAEVHVQQANQQLSRAADYQRKS.RKTLC 239
          .
268 IIICCVILGIVIASTV 283
    |||  .::|: | | :
240 IIILNLVIGVAIISLI 255
```

What do you need to know about this output?

- The alignment uses the one-letter amino acid code; memorize it!
- The lines indicate amino acid identities.
- The dots indicate similarity; this is sometimes less informative than identity.
- The gap weight and gap length penalties are parameters you can adjust to optimize the alignment you are performing. For example, if you believe that two proteins are distantly related but they do not align well, try reducing the gap penalty to see if a less stringent alignment criterion yields a better alignment (at the cost of adding more gaps).

GCG Documentation on BestFit Program (> genhelp bestfit)

BestFit makes an optimal alignment of the best segment of similarity between two sequences. Optimal alignments are found by inserting gaps to maximize the number of matches using the local homology algorithm of Smith and Waterman (Chapter 3).

Description. BestFit inserts gaps to obtain the optimal alignment of the best region of similarity between two sequences and then displays the alignment in a format similar to the output from GAP. The sequences can be of very different lengths and have only a small segment of similarity between them. You could take a short RNA sequence, for example, and run it against a whole mitochondrial genome.

Searching for Similarity. BestFit is the most powerful method in the Wisconsin Package for identifying the best region of similarity between two sequences whose relationship is unknown.

Example of How to Use GAP Program

```
> gap humsyn1a.pep humsyn7.pep

Gap uses the algorithm of Needleman and Wunsch to find the alignment
of two complete sequences that maximizes the number of matches and
minimizes the number of gaps.
```

```
          Begin (* 1 *) ?
        End (*   289 *) ?

          Begin (* 1 *) ?
        End (*   261 *) ?

What is the gap creation penalty (* 12 *) ?

What is the gap extension penalty (* 4 *) ?

What should I call the paired output display file (* humsyn1a.pair *) ?

Aligning .............-.
Aligning .............-..

        Gaps:      2
     Quality:    169
Quality Ratio:  0.648
 % Similarity: 35.659
      Length:    292
```

Next, view the resulting pairwise alignment.

```
> cat humsyn1a.pair
 GAP of: humsyn1a.pep  check: 5551  from: 1  to: 289

LOCUS       HUMSYN1A     2088 bp    mRNA        PRI 23-NOV-1994
DEFINITION  Human syntaxin 1A mRNA, complete cds.
ACCESSION   L37792
NID         g577487
KEYWORDS    syntaxin 1A.
SOURCE      Homo sapiens male adult brain cDNA to mRNA . . .

 to: humsyn7.pep  check: 338  from: 1  to: 261

 Symbol comparison table: /prog/GCG/gcgcore/data/rundata/blosum62.cmp
 CompCheck: 6430

        Gap Weight:    12     Average Match:  2.912
     Length Weight:     4  Average Mismatch: -2.003

           Quality:   169            Length:    292
             Ratio: 0.648              Gaps:      2

 Percent Similarity: 35.659 Percent Identity: 23.643

        Match display thresholds for the alignment(s):
            | = IDENTITY
            : =   2
            . =   1

humsyn1a.pep x humsyn7.pep October 28, 1998 20:08  ..

          .         .         .         .         .
     1 MKDRTQELRTAKDSDDDDVAVTVDRDRFMDEFFEQVEEIRGFIDKIAEN 50
                                  |         :      :|. |
     1 ..........................MSYTPGVGGDPTQLAQRISSN 21
               .         .         .         .         .
    51 VEEVKRKHSAILASPNPDEKTKEELEELMSDIKKTANKVRSKLKSIEQSI 100
       ::.: .    |  . |    | ::  ||   :..         |  :. |
    22 IQKITQCSVEIQRTLN.QLGTPQDSPELRQQLQQKQQYTNQLAKETDKYI 70
               .         .         .         .         .
```

```
101 EQEEGLNRSSADLRIRKTQHSTLSRKFVEVMSEYNATQSDYRERCK...G 147
    .:   |  . .: | || |   |  .    :. :   |   ||  |
 71 KEFGSLPTTPSEQRQRKIQKDRLVAECTTSLTNFQKVQRQAAEREKEFVA 120
            .         .         .         .         .
148 RIQRQLEITGRTTTSEELEDMLESGNPAIFASGIIMDSSISKQALSEIET 197
    |:.    :.|        |  | |          : |  |..  |  |
121 RVRASSRVSGSFPEDSSKERNLVSWESQTQPQVQVQDEEITEDDLRLIHE 170
            .         .         .         .         .
198 RHSEIIKLENSIRELHDMFMDMAMLVESQGEMIDRIEYNVEHAVDYVERA 247
    | | | .||  | ::.:.| |: |::  ||:.|| || |||.|  :|:.|
171 RESSIRQLEADIMDINEIFKDLGMMIHEQGDVIDSIEANVENAEVHVQQA 220
            .         .         .         .
248 VSDTKKAVKYQSKARRKKIMIIICCVILGIVIASTVGGIFA* 289
        :|  || |.|:   .||:  ||   :|.  : |:
221 NQQLSRAADYQRKSRKTLCIIILNLVIGVAIISLIIWGLNH. 261
```

Is Homology of Two Proteins Statistically Significant?

For local alignments, the most rigorous approach to determining whether two nucleic acid or protein sequences are homologous is to measure the expectation value (as described in Chapter 4). For global alignments, the statistical significance is harder to assess. One approach is to perform a randomization test (see Figure 3.30). This can be performed in GCG as follows.

Type "> gap humsyn1a.pep humsyn7.pep -ran = 50."

The "-ran=50" command asks the program to compare the proteins humsyn1a.pep and humsyn7.pep to obtain a quality score. Then the program takes the amino acid sequence of humsyn7.pep and randomly shuffles it. This randomized humsyn7.pep is compared to humsyn1a.pep, and a quality score is generated. Fifty such randomized humsyn7.pep proteins are compared to humsyn1a.pep, one at a time. The program provides the usual output of GAP, with the addition of the mean and standard deviation of the quality scores of 50 randomizations. For example, the output now includes:

```
          Gap Weight:     12          Average Match:  2.912
       Length Weight:      4       Average Mismatch: -2.003

             Quality:    194                 Length:    216
               Ratio:  0.911                   Gaps:      2
 Percent Similarity: 39.151       Percent Identity: 26.887

 Average quality based on 50 randomizations: 35.2 +/- 4.0
```

These data of the average quality based on 50 randomizations let you answer the key question: When I get a quality score from the comparison of my two proteins ($Q = 194$ in this case), is that score much better than if I had just randomly shuffled one of the proteins 50 times ($Q = 35.2 \pm 4.0$)? The answer is yes, because the Z score for the authentic comparison is more than three standard deviations greater than the randomized score. For this example, $Z = (194 - 35.2)/4.0 = 39.7$. If Z less than 3, your two proteins may not be significantly related.

GAP (Global Alignment) versus BestFit (Local Alignment; BLAST-like)

Note that while BestFit and GAP are similar—each allows you to compare two sequences—they do have distinct differences. BestFit is optimized for *local* alignments, finding the best region of homology between two proteins (or DNA sequences). GAP is optimized for *global* alignments, scanning the entire sequences.

In the above example of two syntaxins, BestFit showed a better percent amino acid identity (27% over a length of 216 residues), while GAP showed only 24% identity but over a longer span of 292 residues.

GCG Documentation on GAP Program (> genhelp gap)

GAP uses the algorithm of Needleman and Wunsch to find the alignment of two complete sequences that maximizes the number of matches and minimizes the number of gaps.

Description

GAP considers all possible alignments and gap positions and creates the alignment with the largest number of matched bases and the fewest gaps. You provide a gap creation penalty and a gap extension penalty in units of matched bases. In other words, GAP must make a profit of gap creation penalty number of matches for each gap it inserts. If you choose a gap extension penalty greater than zero, GAP must, in addition, make a profit for each gap inserted of the length of the gap times the gap extension penalty. GAP uses the alignment method of Needleman and Wunsch [*J. Mol. Biol.* 48; 443-453 (1970)].

Example of How to Use TFastA Program

```
> tfasta humsyn1a.pep humsyn7.seq

TFastA does a Pearson and Lipman search for similarity between a query
peptide sequence and any group of nucleotide sequences. TFastA translates
the nucleotide sequences in all six reading frames before performing the
comparison. It is designed to answer the question, "What implied peptide
sequences in a nucleotide sequence database are similar to my peptide
sequence?"

               Begin (* 1 *) ?
             End (*   289 *) ?

 What word size (* 2 *) ?

 Don't show scores whose E() value exceeds: (* 10.0 *):

 What should I call the output file (* humsyn1a.tfasta *) ?

          1 Sequences       1,673 aa searched
/user/pevsner/pevsner/syntaxin.
dir/humsyn7.seq

(Peptide) TFASTA of: humsyn1a.pep from: 1 to: 289 October 28, 1998 22:02
TO: humsyn7.seq Sequences:           1 Symbols:    1,673  Word Size: 2

Searching all six frames.
Scoring matrix: GenRunData:blosum50.cmp
Variable pamfactor used
Gap creation penalty: 16       Gap extension penalty: 4

The best scores are:                    frame init1  initn  opt
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq
!                                         (3)    171    189  268
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq
!                                         (2)     44     44   49
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq
!                                         (4)     38     38   41
```

```
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq
!                                      (1)     29     29   35
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq
!                                      (5)     29     29   29
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq
!                                      (6)     26     26   28
The list contains 6 entries.
How many alignments would you like to see (* 6 *) ?

Aligning...
 CPU time used:
       Database scan: 0:00: 0.1
Post-scan processing: 0:00: 0.1
      Total CPU time: 0:00: 0.3
 Output File: humsyn1a.tfasta

> cat humsyn1a.tfasta
!!SEQUENCE_LIST 1.0

(Peptide) TFASTA of: humsyn1a.pep from: 1 to: 289 October 28, 1998 22:02

 TO: humsyn7.seq Sequences:          1 Symbols:    1,673    Word Size: 2

 Searching all six frames.
 Scoring matrix: GenRunData:blosum50.cmp
 Variable pamfactor used
 Gap creation penalty: 16      Gap extension penalty: 4

The best scores are:                    frame initl  initn  opt..

/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq     Begin: 201  End: 917
!                                      (3)    171   189   268
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq     Begin: 557  End: 598
!                                      (2)     44    44    49
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq     Begin: 831  End: 920 Strand: -
!                                      (4)     38    38    41
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq     Begin: 310  End: 420
!                                      (1)     29    29    35
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq     Begin: 1619  End: 1645 Strand: -
!                                      (5)     29    29    29
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq     Begin: 1060  End: 1119 Strand: -
!                                      (6)     26    26    28
\\End of List
humsyn1a.pep
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq

SCORES  Frame:  (3)  Init1:   171 Initn:   189 Opt:   268
  25.3% identity  in  241 aa overlap

               20        30        40        50        60         70
humsyn1a.pep DDDDVAVTVDRDRFMDEFFEQVEEIRGFIDKIAENVEEVKRKHSAILASPNPDEKTKEEL
                                          :|: |:::: :    |  : | :  | ::
humsyn7      LHSIXXNXRIRTSTMSYTPGVGGDPTQLAQRISSNIQKITQCSVEIQRTLN-QLGTPQDS
               40        50        60        70        80         90

               80        90       100       110       120        130
humsyn1a.pep EELMSDIKKTANKVRSKLKSIEQSIEQEEGLNRSSADLRIRKTQHSTLSRKFVEVMSEYN
              || :::::  : : :  | :: |::  :|  : :: | || |:: |  : :  :::::
humsyn7      PELRQQLQQKQQYTNQLAKETDKYIKEFGSLPTTPSEQRQRKIQKDRLVAECTTSLTNFQ
               100       110       120       130       140        150

               140          150       160       170       180        190
humsyn1a.pep ATQSDYRERCK---GRIQRQLEITGRTTTSEELEDMLESGNPAIFASGIIMDSSISKQAL
              :| : ||  |   :|:: : :::|    :   |  | | :     :  ::|  |::: |
humsyn7      KVQRQAAEREKEFVARVRASSRVSGSFPEDSSKERNLVSWESQTQPQVQVQDEEITEDDL
               160       170       180       190       200        210
```

```
                    200       210       220       230       240       250
humsyn1a.pep SEIETRHSEIIKLENSIRELHDMFMDMAMLVESQGEMIDRIEYNVEHAVDYVERAVSDTK
               |: |:| | :|| :| ::::: |::|::: ||::|| || |||:|  :|::| :: :
humsyn7      RLIHERESSIRQLEADIMDINEIFKDLGMMIHEQGDVIDSIEANVENAEVHVQQANQQLS
               220       230       240       250       260       270

                    260       270       280       289
humsyn1a.pep KAVKYQSKARRKKIMIIICCVILGIVIASTVGGIFAX
             :|: || |: || : |||  :::|::| | :
humsyn7      RAADYQRKS-RKTLCIIILNLVIGVAIISLIIWGLNHXSYKGAHCRTTLSKLCRKIPVIM
               280       290       300           310    320       330

humsyn7      FFXLLFXSYCIKDGSHTLLFLLGGLGGFLWIKSDIFXYXKIFXMSLLTXLPWSSIXXLLS
               340       350       360           370    380       390

humsyn1a.pep
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq

SCORES  Frame: (2) Init1:    44 Initn:   44 Opt:   49
  50.0% identity in 14 aa overlap

                     30        40        50        60        70        80
humsyn1a.pep TVDRDRFMDEFFEQVEEIRGFIDKIAENVEEVKRKHSAILASPNPDEKTKEELEELMSDI
                                           :::||:| ||::||
humsyn7      EGPEAGCXARERVCCSSKSQFQSVWQFSXGQLQRKESCILGKPNSTSSAGAGXRNYRGXP
               160       170       180       190       200       210

                     90        100       110       120       130       140
humsyn1a.pep KKTANKVRSKLKSIEQSIEQEEGLNRSSADLRIRKTQHSTLSRKFVEVMSEYNATQSDYR

humsyn7      PSYSXERIFYQATXSXYYGYXXNIXRFGNDDSXTRRCNRXHRSQCGKCRGARSASKSAAV
               220       230       240       250       260       270
humsyn1a.pep
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq

SCORES  Frame: (4) Init1:    38 Initn:   38 Opt:   41
  26.7% identity in 30 aa overlap

                    220       230       240       250       260       270
humsyn1a.pep HDMFMDMAMLVESQGEMIDRIEYNVEHAVDYVERAVSDTKKAVKYQSKARRKKIMIIICC
                                          | |    :::  || ::    : : :||||
humsyn7      KKHDYRNLPTXFRQCSATVCSFITSVVQSPYDETDNRNSNDKVKNDDAQGFSGFALIICC
                     230       240       250       260       270       280
                    280      289

humsyn1a.pep VILGIVIASTVGGIFAX

humsyn7      PXQLLICLLNVHLCIFHIGFYAIYYISLFMNHHSQIFKYFINIHNISFKLPDRRFSLMNK
                     290       300       310       320       330       340

humsyn1a.pep
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq

SCORES  Frame: (1) Init1:    29 Initn:   29 Opt:   35
  35.1% identity in 37 aa overlap

                   80        90        100       110       120       130
humsyn1a.pep ELMSDIKKTANKVRSKLKSIEQSIEQEEGLNRSSADL-RIRKTQHSTLSRKFVEVMSEYN
                                          ||||: |  : | | |||:   :      |
humsyn7      RRSHSVLWKYKELXINLEHLKIHLNXGNSCNRSSSILTSLPKKQISTLKSLDLCPPPPVN
                     80        90        100       110       120       130
```

```
              140       150        160        170        180       190
humsyn1a.pep ATQSDYRERCKGRIQRQLEITGRTTTSEELEDMLESGNPAIFASGIIMDSSISKQALSEI
             :::: ||:
humsyn7      SVKGKYRRIAXWQSAQHHXQTSRRSRGRLLSERKSLLLEXEPVPECLAVFLRTAPKKGIL
                 140       150        160        170        180       190

humsyn1a.pep
/user/pevsner/pevsner/syntaxin.dir/humsyn7.seq

! CPU time used:
!        Database scan: 0:00:00.1
! Post-scan processing: 0:00:00.1
!       Total CPU time: 0:00:00.3
! Output File: humsyn1a.tfasta
```

When Do You Use TFastA?

- If you obtain a DNA sequence (e.g., from a BLAST search or from the core facility following a yeast two-hybrid result), a TFastA search will tell you its relation to any given protein. If there are frame shifts due to sequencing errors, the TFastA search will show this. Using the GCG map program, errors can be corrected.
- TFastA can be a useful tool for new gene discovery (a process called *database mining*, see Chapter 5). If you are studying a gene family and want to find a new member, go to GenBank and conduct a series of tblastn searches with your proteins of interest against all the databases of expressed sequence tags. Download all DNA database entries relevant to your protein queries into GCG. Then perform TFastA searches to assign each expressed sequence tag (EST) (i.e., DNA sequence) to a protein. If an EST does *not* match any protein in your family, you have identified a novel gene.

TOPIC 5: MULTIPLE SEQUENCE ALIGNMENT

GCG Documentation on PileUp Program (> genhelp pileup)

Function of PileUp

PileUp creates a multiple sequence alignment from a group of related sequences using progressive, pairwise alignments. It can also plot a tree showing the clustering relationships used to create the alignment.

Description of PileUp

PileUp creates a multiple sequence alignment using a simplification of the progressive alignment method of Feng and Doolittle [*J. Mol. Evol.* 25; 351–360 (1987)]. The method used is similar to the method described by Higgins and Sharp [*CABIOS* 5; 151–153 (1989)].

The multiple alignment procedure begins with the pairwise alignment of the two most similar sequences, producing a cluster of two aligned sequences. This cluster can then be aligned to the next most related sequence or cluster of aligned sequences. Two clusters of sequences can be aligned by a simple extension of the pairwise alignment of two individual sequences. The final alignment is achieved by a series of progressive, pairwise alignments that include increasingly dissimilar

sequences and clusters, until all sequences have been included in the final pairwise alignment.

Before alignment, the sequences are first clustered by similarity to produce a dendrogram, or tree representation of clustering relationships. It is this dendrogram that directs the order of the subsequent pairwise alignments. PileUp can plot this dendrogram so that you can see the order of the pairwise alignments that created the final alignment.

As a general rule, PileUp can align up to 500 sequences, with any single sequence in the final alignment restricted to a maximum length of 7000 characters (including gap characters inserted into the sequence by PileUp to create the alignment). However, if you include long sequences in the alignment, the number of sequences PileUp can align decreases.

Example of PileUp

There are two easy ways to run PileUp:

1. On the command line, type "> PileUp *.pep." All files with the suffix .pep will be made into the pileup.
2. If your directory has many files, you can create a special file called "anyname" that includes the specific proteins (or DNA sequences) you want to have included:

```
> cat> anyname
a.pep
b.pep
c.pep
[^Y]
> pileup @anyname
```

Here is a sample output:

```
> pileup @x

PileUp creates a multiple sequence alignment from a group of related
sequences using progressive, pairwise alignments. It can also plot a
tree showing the clustering relationships used to create the alignment.

   1    humsyn1a.pep    289  aa
   2     ratsyn2.pep    291  aa
   3     humsyn3.pep    290  aa
   4     humsyn4.pep    298  aa
   5     humsyn5.pep    302  aa
   6     ratsyn6.pep    255  aa
   7     humsyn7.pep    261  aa
   8   humsyn16b.pep    324  aa

What is the gap creation penalty (* 12 *)  ?  8

What is the gap extension penalty (* 4 *)  ?  2

This program can display the clustering relationships graphically.
Do you want to:

    A) Plot to a FIGURE file called "pileup.figure"
    B) Plot graphics on HP7550 attached to /dev/tty15
    C) Suppress the plot
```

```
Please choose one (* A *):

The minimum density for a one-page plot is 6.7 sequences/100 platen units.
What density do you want (* 6.7 *) ?

What should I call the output file name (* x.msf *) ?

Determining pairwise similarity scores...

  1   x     2       3.15
  1   x     3       3.02
  1   x     4       2.09
          ...
  5   x     8       0.27
  6   x     7       0.15
  6   x     8       0.29
  7   x     8       0.79

Aligning...

  1     ..............-..
  2     ..............-.
        ..............-..
  3     ..............-.
        ..............-..
  4     ..............-..
  5     ..............-..
  6     ..............-..
  7     ............-..
```

If you specify option A, you will get a multiple sequence alignment of your proteins: For example:

```
> cat x.msf
!!AA_MULTIPLE_ALIGNMENT 1.0
PileUp    of: @x

 Symbol  comparison   table: GenRunData:blosum62.cmp   CompCheck:  6430
                     GapWeight: 8
              GapLengthWeight: 2

 x.msf   MSF: 363    Type:  P  March 9, 1998  13:42  Check:  7155 ..

 Name:  humsyn1a             Len:   363  Check:  3616  Weight:   1.00
 Name:  ratsyn2              Len:   363  Check:  1439  Weight:   1.00
 Name:  humsyn3              Len:   363  Check:  7304  Weight:   1.00
 Name:  humsyn4              Len:   363  Check:  9241  Weight:   1.00
 Name:  humsyn7              Len:   363  Check:  8384  Weight:   1.00
 Name:  humsyn16b            Len:   363  Check:  2604  Weight:   1.00
 Name:  humsyn5              Len:   363  Check:  2749  Weight:   1.00
 Name:  ratsyn6              Len:   363  Check:  1818  Weight:   1.00

//
          1                                                      50
 humsyn1a  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~MKDRTQELR
  ratsyn2  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~MRDRLPDLT
  humsyn3  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~MKDRLEQLK
  humsyn4  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~MRDRTHELR
  humsyn7  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~
humsyn16b  MATRRLTDAF  LLLRNNSIQN  RQLLAEQVSS  HITSSPLHSR  SIAAELDELA
  humsyn5  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~M  SCRDRTQEFL
  ratsyn6  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~
```

```
           51                                                        100
 humsyn1a  TAK.DSDDDD  .D.VAVTVDR  DRFM.....D  EFFEQVEEIR  GFIDKIAENV
  ratsyn2  ACR.KSDDGD  .NAVIITVEK  DHFM.....D  AFFHQVEEIR  SSIARIAQHV
  humsyn3  AKQLTQDDDT  .DAVEIAIDN  TAFM.....D  EFFSEIEETR  LNIDKISEHV
  humsyn4  QGDDSSDEED  KERVALVVHP  GTARLGSPDE  EFFHKVRTIR  QTIVKLGNKV
  humsyn7  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~MSYTP  GVGGDPTQLA  QRISSNIQKI
humsyn16b  DDRMALVSGI  SLDPEAAIGV  TKRPPPKWVD  GV....DEIQ  YDVGRIKQKM
  humsyn5  SACKSLQTRQ  N...GIQTNK  PALRAVRQRS  EFTLMAKRIG  KDLSNTFAKL
  ratsyn6  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~~~~  ~~~~~~~MSM

           101                                                       150
 humsyn1a  EE..VK..RK  HSAIL...AS  PNPDEK..TK  EELEELMSDI  KKTANKVRSK
  ratsyn2  ED..VK..KN  HSIIL...SA  PNPEGK..IK  EELEDLNKEI  KKTANRIRGK
  humsyn3  EE..AK..KL  YSIIL...SA  PIPEPK..TK  DDLEQLTTEI  KKRANNVRNK
  humsyn4  QE..LE..KQ  QVTIL...AT  PLPEES..MK  QELQNLRDEI  KQLGREIRLQ
  humsyn7  TQ..CS..VE  IQRTLNQLGT  PQDSPE..LR  QQLQQKQQYT  NQLAKETDKY
humsyn16b  KE..SA..SL  HDKHLNR.PT  LDDSSE..EE  HAIEITTQEI  TQLFHRCQRA
  humsyn5  EK..LTILAK  RKSLFDDKAV  EIEELTYIIK  QDINSLNKQI  AQLQDFVRAK
  ratsyn6  EDPFFVVKGE  VQKAVNTAQG  LFQRWTELLQ  GPSAATREEI  DWTTNELRNN
```

If you specify option B, you will get a tree (see below).

Topic 6: Phylogeny

What Is a Phylogenetic Tree?

A phylogenetic tree is a graphical representation of the evolutionary history of a protein family (or other family such as a gene family) (see Chapter 11). Phylogeny reconstructs the branching pattern of a group of proteins. The topology of a tree shows the history (ancestors) of the protein, in some cases over time.

Why Are Trees Useful?

A phylogenetic tree provides a visual demonstration of how a group of nucleic acid or protein sequences are related. Suppose you are studying a particular kinase cloned from a mouse cDNA library. A tree can help answer the following questions:

- Does your kinase have closely related homologs? For example, if you knock out its gene, are there closely related kinases? If you alter the kinase activity in cells with a drug, are there other kinases likely to be affected as well?
- How should you name it?
- Has it been relatively conserved throughout evolution or has its sequence changed rapidly? This may give clues as to its function.
- What individual amino acid residues are critical for the function of a kinase?

How Do I Make a Tree?

To make a tree, begin by doing a PileUp of your group of protein or nucleic acid sequences. The output of PileUp can be a multiple sequence alignment, with a suffix pileup.msf (multiple sequence format). Alternatively, you can specify the output to be a simple tree by choosing output option B.

We described how to make a tree in Chapter 11. A key tree-making programs in GCG is distances, which makes a matrix of pairwise distances between the sequences in a multiple sequence alignment. Diverge is another program that measures the number of synonymous and nonsynonymous substitutions per site of two or more aligned protein-coding regions and can output matrices of these values. GrowTree reconstructs a tree from a distance matrix or a matrix of synonymous or nonsynonymous substitutions.

In Chapter 11 we described PAUP (Phylogeny Analysis Using Parsimony), a superb tree-building program. PAUP has also been incorporated into GCG. PAUPSearch reconstructs phylogenetic trees from a multiple sequence alignment using parsimony, distance or maximum-likelihood criteria; PAUPDisplay can manipulate and display the trees that are output by PAUPSearch and can also plot the trees that are output by GrowTree.

Topic 7: Sequence Analysis

Prime	Designs oligonucleotide primers
Map	Makes a map of a DNA sequence; shows restriction enzymes
Map six	Makes a map showing restriction enzyme six-cutters
Motifs x.pep	Scans the PROSITE dictionary for motifs and domains present in protein x

GCG Documentation on Prime Program (> genhelp prime)

Function

Prime selects oligonucleotide primers for a template DNA sequence. The primers may be useful for the polymerase chain reaction (PCR) or for DNA sequencing. You can allow Prime to choose primers from the whole template or limit the choices to a particular set of primers listed in a file.

Description

Prime analyzes a template DNA sequence and chooses primer pairs for the PCR and primers for DNA sequencing. For PCR primer pair selection, you can choose a target range of the template sequence to be amplified. For DNA sequencing primers, you can specify positions on the template that must be included in the sequencing.

GCG Documentation on Map Program (> genhelp map)

Function

Map maps a DNA sequence and displays both strands of the mapped sequence with restriction enzyme cut points above the sequence and protein translations below. Map can also create a peptide map of an amino acid sequence.

Description

Map displays a sequence that is being assembled or analyzed intensively. Map asks you to select the enzymes whose restriction sites should be marked individually by

typing their names. If you do not answer this question, Map selects a representative isoschizomer from all of the commercially available enzymes. You can choose to have your sequence translated in any or all of the six possible translation frames. You can also choose to have only the open reading frames translated.

After running Map, you may create a new sequence file with the peptide sequence from any frame of DNA translation by using the ExtractPeptide program with the Map output file.

Example of How to Use Map Program

Here we input a DNA sequence (human syntaxin 7) and specify that we want the output to show only enzymes that are six-cutters (158 enzymes including *Eco*RI, *Pvu*II, *Bam*HI, etc.). In all, the program can analyze sites for over 500 restriction enzymes. Note that you can specify an output display ranging from no predicted proteins to all six potential reading frames.

```
> map humsyn7.seq -six

Map maps a DNA sequence and displays both strands of the mapped sequence
with restriction enzyme cut points above the sequence and protein
translations below. Map can also create a peptide map of an amino acid
sequence.

                Begin  (* 1 *) ?
              End  (*  1673 *) ?

Select the enzymes:  Type nothing or "*" to get all enzymes. Type "?"
for help on which enzymes are available and how to select them.

                                      Enzyme(* * *):

What protein translations do you want:

     a) frame 1   b) frame 2   c) frame 3
     d) frame 4   e) frame 5   f) frame 6

     t)hree forward frames   s)ix frames   o)pen frames only

     n)o protein translation   q) uit

Please select (capitalize for 3-letter) (* t *):

What should I call the output file (* humsyn7.map *) ?

Mapping ..............................

Writing ............. ..
MAP complete with:

  Sequence Length:       1,673
   Enzymes Chosen:         158
   Cutsites found:          94
         CPU time:       00.28

   Output file(s):   humsyn7.map

> cat humsyn7.map
(Linear)  (MinSite=6) MAP of: humsyn7.seq check: 612 from: 1 to: 1673

With 158 enzymes: *
```

```
                          October 29, 1998 12:18   ..

                                                 MmeI
                                 Bpu10I    TaqII
                    MspA1I       MspA1I    BanII
                     PvuII        PvuII  Bsp1286I                 PshAI
                       |            |        |                      |
      CGCCACCGCCCATCAGCTGAGAATTGCAGCTGAGGGCTCCGGGGTAGGTGGGTGACGGCG
    1 ---------+---------+---------+---------+---------+---------+ 60
      GCGGTGGCGGGTAGTCGACTCTTAACGTCGACTCCCGAGGCCCCATCCACCCACTGCCGC
a     R  H  R  P  S  A  E  N  C  S  *  G  L  R  G  R  W  V  T  A   -
b      A  T  A  H  Q  L  R  I  A  A  E  G  S  G  V  G  G  *  R  R  -
c       P  P  P  I  S  *  E  L  Q  L  R  A  P  G  *  V  G  D  G  G -
```

GCG Documentation on Motifs Program (> genhelp motifs)

Function

Motifs looks for sequence motifs by searching through proteins for the patterns defined in the PROSITE Dictionary of Protein Sites and Patterns (see Chapter 8). Motifs can display an abstract of the current literature on each of the motifs it finds.

Description

Motifs looks for protein sequence motifs by checking your protein sequence for every sequence pattern in the PROSITE Dictionary. Motifs can recognize the patterns with some of the symbols mismatched, but not with gaps. Currently, Motifs can only search for patterns in protein sequences. There is a very informative abstract on every motif in the PROSITE Dictionary. These abstracts are displayed next to any motif found in your sequence.

Example of How to Use Motifs Program

```
> motifs humsyn1a.pep

Motifs looks for sequence motifs by searching through proteins for the
patterns defined in the PROSITE Dictionary of Protein Sites and
Patterns. Motifs can display an abstract of the current literature on
each of the motifs it finds.

 What should I call the output file (* humsyn1a.motifs *) ?

        humsyn1a.pep len:        289 ............................

             Total finds:          1
           Total length:         289
         Total sequences:          1
          CPU time (sec):      01.63
             Output file: "/user/pevsner/pevsner/syntaxin.dir/humsyn1a.motifs"

> cat humsyn1a.motifs
MOTIFS from: humsyn1a.pep

Mismatches: 0                   October 29, 1998 12:28 ..
       humsyn1a.pep     Check: 5551  Length: 289     ! LOCUS    HUMSYN1A    2088
bp    mRNA               PRI       23-NOV-1994
-------------------------------------------------------------------------------

Epimorphin            Rx3(L,I,V,M)x2(L,I,V,M)Ex2(L,I,V,M)xE(L,I,V,M)x2(L,I,V,M)F
x2(L,I,V,M)x2(L,I,V,M,F)Vx2Q(G,Q)x2(L,I,V,M)(D,N,Q)x(L,I,V,M)(D,E)xN
                      Rx{3}(I)x{2}(L)Ex{2}(I)xE(L)x{2}(M)Fx{2}(M)x{2}(L)Vx{2}Q
(G)x{2}(I)(D)x(I)(E)xN
          198: SEIET   RHSEIIKLENSIRELHDMFMDMAMLVESQGEMIDRIEYN   VEHAV
```

```
******************************
* Epimorphin family signature *
******************************

The following proteins have been shown to be evolutionary related [1,2,3]:

 - Epimorphin, a mammalian mesenchymal protein which plays an essential role in
   epithelial morphogenesis.
 - Syntaxin A (also known as antigen HPC-1) and syntaxin B which are synaptic
   proteins which may be involved in docking of synaptic vesicles at presynaptic
   active zones.
 - Yeast PEP12 which is required for the transport of proteases to the vacuole.
 - Yeast SED5 which is required for the fusion of transport vesicles with the
   Golgi complex.
 - Yeast SSO1 and SSO2 which are required for vesicle fusion with the plasma
   membrane.

The above proteins share the following characteristics: a size ranging from
30 Kd to 40 Kd; a C-terminal extremity which is highly hydrophobic and is
probably involved in anchoring the protein to the membrane; a central, well
conserved region, which seems to be in a coiled-coil conformation.

The pattern specific for this family is based on the most conserved region of
the coiled coil domain.

-Consensus pattern: R-x(3)-[LIVM]-x(2)-[LIVM]-E-x(2)-[LIVM]-x-E-[LIVM]-x(2)-
                    [LIVM]-F-x(2)-[LIVM]-x(2)-[LIVMF]-V-x(2)-Q-[GQ]-x(2)-
                    [LIVM]-[DNQ]-x-[LIVM]-[DE]-x-N
-Sequences known to belong to this class detected by the pattern: ALL.
-Other sequence(s) detected in SWISS-PROT: NONE.
-Last update: October 1993 / First entry.
```

Glossary

The glossaries are on-line at:

- http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html
- http://www.ornl.gov/TechResources/Human_Genome/glossary/
- http://www.genome.gov/page.cfm?pageID=10002096
- http://smart.embl-heidelberg.de/help/smart_glossary.shtml
- http://scop.mrc-lmb.cam.ac.uk/scop/gloss.html

This glossary is combined from five web-based glossaries and each entry is marked accordingly: (1) the National Center for Biotechnology Information (NCBI), (2) the Oak Ridge National Laboratory (ORNL), (3) the talking glossary at the National Human Genome Research Institute (NHGRI), (4) the SMART database, and (5) the protein folds glossary from the Structural Classification of Proteins website (SCOP) (these entries are modified). Additional web-based glossaries are listed in a table at the end of this glossary.

A

Additive genetic effects
When the combined effects of alleles at different loci are equal to the sum of their individual effects. (ORNL)

Adenine (A)
A nitrogenous base, one member of the base pair AT (adenine-thymine). *See also:* base pair. (ORNL)

Algorithm
A fixed procedure embodied in a computer program. (NCBI)

Alignment
(a) The process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology. (NCBI) (b) Representation of a prediction of the amino acids in tertiary structures of homologs that overlay in three dimensions. Alignments held by SMART are mostly based on published observations (see domain annotations for details) but are updated and edited manually. (SMART)

All alpha
A class that has the number of secondary structures in the domain or common core described as 3-, 4-, 5-, 6-, or multihelical. (SCOP)

All beta
A class that includes two major fold groups: sandwiches and barrels. The sandwich folds are made of two β sheets which are usually twisted and packed so their strands are aligned. The barrel fold are made of a single β sheet that twists and coils upon itself so, in most cases, the first strand in the β sheet hydrogen bonds to the last strand. The strand directions in the two opposite sides of a barrel fold are roughly orthogonal. Orthogonal packing of sheets is also seen in a few special cases of sandwich folds. (SCOP)

Allele
(a) Alternative form of a genetic locus; a single allele for each locus is inherited from each parent (e.g., at a locus for eye color the allele might result in blue or brown eyes). (ORNL) (b) One of the variant forms of a gene at a particular locus, or location, on a chromosome. Different alleles produce variation in inherited characteristics such as hair color or blood type. In an individual, one form of the allele (the dominant one) may be expressed more than another form (the recessive one). (NHGRI)

Allogeneic
Variation in alleles among members of the same species. (ORNL)

Alternative splicing
Different ways of combining a gene's exons to make variants of the complete protein. (ORNL)

Amino acid
Any of a class of 20 molecules that are combined to form proteins in living things. The sequence of amino acids in a protein and hence protein function are determined by the genetic code. (ORNL)

Amplification
An increase in the number of copies of a specific DNA fragment; can be in vivo or in vitro. *See also:* cloning. (ORNL)

Animal model
See: model organisms. (ORNL)

Annotation
Adding pertinent information such as gene coded for, amino acid sequence, or other commentary to the database entry of raw sequence of DNA bases. *See also:* bioinformatics. (ORNL)

Anticipation
Each generation of offspring has increased severity of a genetic disorder; e.g., a grandchild may have earlier onset and more severe symptoms than the parent, who had earlier onset than the grandparent. *See also:* additive genetic effects, complex trait. (ORNL)

Antisense
Nucleic acid that has a sequence exactly opposite to an mRNA molecule made by the body; binds to the mRNA molecule to prevent a protein from being made. *See also:* transcription. (ORNL)

Apoptosis
Programmed cell death, the body's normal method of disposing of damaged, unwanted, or unneeded cells. (ORNL)

Array (of hairpins)
An assemble of α helices that cannot be described as a bundle or a folded leaf. (SCOP)

Arrayed library
Individual primary recombinant clones (hosted in phage, cosmid, YAC, or other vector) that are placed in two-dimensional arrays in microtiter dishes. Each primary clone can be identified by the identity of the plate and the clone location (row and column) on that plate. Arrayed libraries of clones can be used for many applications, including screening for a specific gene or genomic region of interest. *See also:* library, genomic library, gene chip technology. (ORNL)

Assembly
Putting sequenced fragments of DNA into their correct chromosomal positions. (ORNL)

Autoradiography
A technique that uses X-ray film to visualize radioactively labeled molecules or fragments of molecules; used in analyzing length and number of DNA fragments after they are separated by gel electrophoresis. (ORNL)

Autosomal dominant
A gene on one of the non-sex chromosomes that is always expressed, even if only one copy is present. The chance of passing the gene to offspring is 50% for each pregnancy. *See also:* autosome, dominant, gene (ORNL)

Autosome
A chromosome not involved in sex determination. The diploid human genome consists of a total of 46 chromosomes: 22 pairs of autosomes and 1 pair of sex chromosomes (the X and Y chromosomes). *See also:* sex chromosome. (ORNL)

B

Backcross
A cross between an animal that is heterozygous for alleles obtained from two parental strains and a second animal from one of those parental strains. Also used to describe the breeding protocol of an outcross followed by a backcross. *See also:* model organisms. (ORNL)

Bacterial artificial chromosome (BAC)
(a) A vector used to clone DNA fragments (100- to 300-kb insert size; average, 150 kb) in *Escherichia coli* cells. Based on naturally occurring F-factor plasmid found in the bacterium *E. coli. See also:* cloning vector. (ORNL) (b) Large segments of DNA, 100,000–200,000 bases, from another species cloned into bacteria. Once the foreign DNA has been cloned into the host bacteria, many copies of it can be made. (NHGRI)

Bacteriophage
See: phage. (ORNL)

Barrel
Structures are usually closed by main-chain hydrogen bonds between the first and last strands of the β sheet; in this case it is defined by the two integer numbers: the number of strand in the β sheet, n, and a measure of the extent to which the strands in the sheet are staggered, the shear number S. (SCOP)

Base
One of the molecules that form DNA and RNA molecules. *See also:* nucleotide, base pair, base sequence. (ORNL)

Base pair (bp)
Two nitrogenous bases (adenine and thymine or guanine and cytosine) held together by weak bonds. Two strands of DNA are held together in the shape of a double helix by the bonds between base pairs. (ORNL)

Base sequence
The order of nucleotide bases in a DNA molecule; determines the structure of proteins encoded by that DNA. (ORNL)

Base sequence analysis
A method, sometimes automated, for determining the base sequence. (ORNL)

Behavioral genetics
The study of genes that may influence behavior. (ORNL)

Beta (β) sheet
Can be antiparallel (i.e., the strand direction in any two adjacent strands are antiparallel), parallel (all strands are parallel to each other), and mixed (there is one strand at least that is parallel to one of its two neighbors and antiparallel to the other). (SCOP)

Bioinformatics
(a) The merger of biotechnology and information technology with the goal of revealing new insights and principles in biology. (NCBI) (b) The science of managing and analyzing biological data using advanced computing techniques. Especially important in analyzing genomic research data. (ORNL)

Bioremediation
The use of biological organisms such as plants or microbes to aid in removing hazardous substances from an area. (ORNL)

Biotechnology
A set of biological techniques developed through basic research and now applied to research and product development. In particular, biotechnology refers to the use by industry of recombinant DNA, cell fusion, and new bioprocessing techniques. (ORNL)

Birth defect
Any harmful trait, physical or biochemical, present at birth, whether a result of a genetic mutation or some other nongenetic factor. *See also:* congenital, gene, mutation, syndrome. (ORNL)

Bit score
(a) The value S' is derived from the raw alignment score S in which the statistical properties of the scoring system used have been taken into account. Because bit scores have been normalized with respect to the scoring system, they can be used to compare alignment scores from different searches. (NCBI) (b) Alignment scores are reported by HMMer and BLAST as bits scores. The likelihood that the query sequence is a bona fide homolog

of the database sequence is compared to the likelihood that the sequence was instead generated by a "random" model. Taking the logarithm (to base 2) of this likelihood ratio gives the bits score. (SMART)

BLAST
(a) Basic Local Alignment Search Tool. A sequence comparison algorithm optimized for speed used to search sequence databases for optimal local alignments to a query. The initial search is done for a word of length W that scores at least T when compared to the query using a substitution matrix. Word hits are then extended in either direction in an attempt to generate an alignment with a score exceeding the threshold of S. The T parameter dictates the speed and sensitivity of the search. For additional details, see one of the BLAST tutorials (Query or BLAST) or the narrative guide to BLAST. (NCBI) (b) A computer program that identifies homologous (similar) genes in different organisms, such as human, fruit fly, or nematode. (ORNL)

BLOSUM
Blocks Substitution Matrix. A substitution matrix in which scores for each position are derived from observations of the frequencies of substitutions in blocks of local alignments in related proteins. Each matrix is tailored to a particular evolutionary distance. In the BLOSUM62 matrix, for example, the alignment from which scores were derived was created using sequences sharing no more than 62% identity. Sequences more identical than 62% are represented by a single sequence in the alignment so as to avoid overweighting closely related family members. (NCBI)

Bundle
An array of α helices each oriented roughly along the same (bundle) axis. It may have twist, left handed if each helix makes a positive angle to the bundle axis or right handed if each helix makes a negative angle to the bundle axis. (SCOP)

C

Cancer
Diseases in which abnormal cells divide and grow unchecked. Cancer can spread from its original site to other parts of the body and can be fatal. *See also:* hereditary cancer, sporadic cancer. (ORNL)

Candidate gene
A gene located in a chromosome region suspected of being involved in a disease. *See also:* positional cloning, protein. (ORNL)

Capillary array
Gel-filled silica capillaries used to separate fragments for DNA sequencing. The small diameter of the capillaries permit the application of higher electric fields, providing high-speed, high-throughput separations that are significantly faster than traditional slab gels. (ORNL)

Carcinogen
Something which causes cancer to occur by causing changes in a cell's DNA. *See also:* mutagen. (ORNL)

Carrier
An individual who possesses an unexpressed, recessive trait. (ORNL)

cDNA library
A collection of DNA sequences that code for genes. The sequences are generated in the laboratory from mRNA sequences. *See also:* messenger RNA. (ORNL)

Cell
The basic unit of any living organism that carries on the biochemical processes of life. *See also:* genome, nucleus. (ORNL)

Centimorgan (cM)
A unit of measure of recombination frequency. One centimorgan is equal to a 1% chance that a marker at one genetic locus will be separated from a marker at a second locus due to crossing over in a single generation. In human beings, one centimorgan is equivalent, on average, to one million base pairs. *See also:* megabase. (ORNL)

Centromere
A specialized chromosome region to which spindle fibers attach during cell division. (ORNL)

Chimera (plural chimaera)
An organism that contains cells or tissues with a different genotype. These can be mutated cells of the host organism or cells from a different organism or species. (ORNL)

Chloroplast chromosome
Circular DNA found in the photosynthesizing organelle (chloroplast) of plants instead of the cell nucleus, where most genetic material is located. (ORNL)

Chromosomal deletion
The loss of part of a chromosome's DNA. (ORNL)

Chromosomal inversion
Chromosome segments that have been turned 180°. The gene sequence for the segment is reversed with respect to the rest of the chromosome. (ORNL)

Chromosome
The self-replicating genetic structure of cells containing the cellular DNA that bears in its nucleotide sequence the linear array of genes. In prokaryotes, chromosomal DNA is circular and the entire genome is carried on one chromosome. Eukaryotic genomes consist of a number of chromosomes whose DNA is associated with different kinds of proteins. (ORNL)

Chromosome painting
Attachment of certain fluorescent dyes to targeted parts of the chromosome. Used as a diagnostic for particular diseases, e.g., types of leukemia. (ORNL)

Chromosome region p
A designation for the short arm of a chromosome. (ORNL)

Chromosome region q
A designation for the long arm of a chromosome. (ORNL)

Clone
An exact copy made of biological material such as a DNA segment (e.g., a gene or other region), a whole cell, or a complete organism. (ORNL)

Clone bank
See: genomic library. (ORNL)

Cloning
Using specialized DNA technology to produce multiple, exact copies of a single gene or other segment of DNA to obtain enough material for further study. This process, used by researchers in the Human Genome Project, is referred to as cloning DNA. The resulting cloned (copied) collections of DNA molecules are called clone libraries. A second type of cloning exploits the natural process of cell division to make many copies of an entire cell. The genetic makeup of these cloned cells, called a cell line, is identical to the original cell. A third type of cloning produces complete, genetically identical animals such as the famous Scottish sheep, Dolly. *See also:* cloning vector. (ORNL)

Cloning vector
DNA molecule originating from a virus, a plasmid, or the cell of a higher organism into which another DNA fragment of appropriate size can be integrated without loss of the vector's capacity for self-replication; vectors introduce foreign DNA into host cells, where the DNA can be reproduced in large quantities. Examples are plasmids, cosmids, and yeast artificial chromosomes; vectors are often recombinant molecules containing DNA sequences from several sources. (ORNL)

Closed, Partly Opened, and Opened
For all-alpha structures, the extent to which the hydrophobic core is screened by the comprising α helices. *Opened* means that there is space for at least one more helix to be easily attached to the core. (SCOP)

Code
See: genetic code. (ORNL)

Codominance
Situation in which two different alleles for a genetic trait are both expressed. *See also:* autosomal dominant, recessive gene. (ORNL)

Codon
See: genetic code. (ORNL)

Coisogenic or congenic
Nearly identical strains of an organism; they vary at only a single locus. (ORNL)

Comparative genomics
The study of human genetics by comparisons with model organisms such as mice, the fruit fly, and the bacterium *Escherichia coli.* (ORNL)

Complementary DNA (cDNA)
DNA that is synthesized in the laboratory from a messenger RNA template. (ORNL)

Complementary sequence
Nucleic acid-base sequence that can form a double-stranded structure with another DNA fragment by following base-pairing rules (A pairs with T and C with G). The complementary sequence to GTAC, for example, is CATG. (ORNL)

Complex trait
Trait that has a genetic component that does not follow strict Mendelian inheritance. May involve the interaction of two or more genes or gene–environment interactions. *See also:* Mendelian inheritance, additive genetic effects. (ORNL)

Computational biology
See: bioinformatics. (ORNL)

Confidentiality
In genetics, the expectation that genetic material and the information gained from testing that material will not be available without the donor's consent. (ORNL)

Congenital
Any trait present at birth, whether the result of a genetic or nongenetic factor. *See also:* birth defect. (ORNL)

Conservation
Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physicochemical properties of the original residue. (NCBI)

Conserved sequence
A base sequence in a DNA molecule (or an amino acid sequence in a protein) that has remained essentially unchanged throughout evolution. (ORNL)

Contig
Group of cloned (copied) pieces of DNA representing overlapping regions of a particular chromosome. (ORNL)

Contig map
A map depicting the relative order of a linked library of overlapping clones representing a complete chromosomal segment. (ORNL)

Cosmid
Artificially constructed cloning vector containing the *cos* gene of phage lambda. Cosmids can be packaged in lambda phage particles for infection into *Escherichia coli;* this permits cloning of larger DNA fragments (up to 45 kb) than can be introduced into bacterial hosts in plasmid vectors. (ORNL)

Crossing over
The breaking during meiosis of one maternal and one paternal chromosome, the exchange of corresponding sections of DNA, and the rejoining of the chromosomes. This process can result in an exchange of alleles between chromosomes. *See also:* recombination. (ORNL)

Crossover
Connection that links secondary structures at the opposite ends of the structural core and goes across the surface of the domain. (SCOP)

Cytogenetics
The study of the physical appearance of chromosomes. *See also:* karyotype. (ORNL)

Cytological band
An area of the chromosome that stains differently from areas around it. *See also:* cytological map. (ORNL)

Cytological map
A type of chromosome map whereby genes are located on the basis of cytological findings obtained with the aid of chromosome mutations. (ORNL)

Cytoplasmic trait
A genetic characteristic in which the genes are found outside the nucleus, in chloroplasts or mitochondria. Results in offspring inheriting genetic material from only one parent. (ORNL)

Cytoplasmic (uniparental) inheritance
See: cytoplasmic trait. (ORNL)

Cytosine (C)
A nitrogenous base, one member of the base pair GC (guanine and cytosine) in DNA. *See also:* base pair, nucleotide. (ORNL)

D

Data warehouse
A collection of databases, data tables, and mechanisms to access the data on a single subject. (ORNL)

Deletion
A loss of part of the DNA from a chromosome; can lead to a disease or abnormality. *See also:* chromosome, mutation. (ORNL)

Deletion map
A description of a specific chromosome that uses defined mutations–specific deleted areas in the genome–as "biochemical signposts," or markers for specific areas. (ORNL)

Deoxyribonucleotide
See: nucleotide. (ORNL)

Deoxyribose
A type of sugar that is one component of DNA (deoxyribonucleic acid). (ORNL)

Diploid
A full set of genetic material consisting of paired chromosomes, one from each parental set. Most animal cells except the gametes have a diploid set of chromosomes. The diploid human genome has 46 chromosomes. *See also:* haploid. (ORNL)

Directed evolution
A laboratory process used on isolated molecules or microbes to cause mutations and identify subsequent adaptations to novel environments. (ORNL)

Directed mutagenesis
Alteration of DNA at a specific site and its reinsertion into an organism to study any effects of the change. (ORNL)

Directed sequencing
Successively sequencing DNA from adjacent stretches of chromosome. (ORNL)

Disease-associated genes
Alleles carrying particular DNA sequences associated with the presence of disease. (ORNL)

DNA (deoxyribonucleic acid)
The molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between base pairs of nucleotides. The four nucleotides in DNA contain the bases adenine (A), guanine (G), cytosine (C), and thymine (T). In nature, base pairs form only between A and T and between G and C; thus the base sequence of each single strand can be deduced from that of its partner. (ORNL)

DNA bank
A service that stores DNA extracted from blood samples or other human tissue. (ORNL)

DNA probe
See: probe. (ORNL)

DNA repair genes
Genes encoding proteins that correct errors in DNA sequencing. (ORNL)

DNA replication
The use of existing DNA as a template for the synthesis of new DNA strands. In humans and other eukaryotes, replication occurs in the cell nucleus. (ORNL)

DNA sequence
The relative order of base pairs, whether in a DNA fragment, gene, chromosome, or an entire genome. *See also:* base sequence analysis. (ORNL)

Domain
(a) A discrete portion of a protein assumed to fold independently of the rest of the protein and possessing its own function. (NCBI) (b) A discrete portion of a protein with its own function. The combination of domains in a single protein determines its overall function. (ORNL) (c) Conserved structural entities with distinctive secondary structure content and an hydrophobic core. In small disulfide-rich and Zn^{2+}-binding or Ca^{2+}-binding domains, the hydrophobic core may be provided by cystines and metal ions, respectively. Homologous domains with common functions usually show sequence similarities. (SMART)

Domain composition
Proteins with the same domain composition have at least one copy of each of the domains of the query. (SMART)

Domain organization
Proteins having all the domains as the query in the same order. (Additional domains are allowed.) (SMART)

Dominant
An allele that is almost always expressed, even if only one copy is present. *See also:* gene, genome. (ORNL)

Double helix
The twisted-ladder shape that two linear strands of DNA assume when complementary nucleotides on opposing strands bond together. (ORNL)

Draft sequence
The sequence generated by the Human Genome Project that, while incomplete, offers a virtual road map to an estimated 95% of all human genes. Draft sequence data are mostly in the form of 10,000 bp-sized fragments whose approximate chromosomal locations are known. *See also:* sequencing, finished DNA sequence, working draft DNA sequence. (ORNL)

DUST
A program for filtering low-complexity regions from nucleic acid sequences. (NCBI)

E

E value
(a) Expectation value. The number of different alignments with scores equivalent to or better than *S* that are expected to occur in a database search by chance. The lower the *E* value, the more significant the score. (NCBI) (b) This represents the number of sequences with a score greater than or equal to *X* expected absolutely by chance. The *E* value connects the score (*X*) of an alignment between a user-supplied sequence and a database sequence, generated by any algorithm, with how many alignments with similar or greater scores that would be expected from a search of a random-sequence database of equivalent size. Since version 2.0, *E* values are calculated using hidden Markov models, leading to more accurate estimates than before. (SMART)

Electrophoresis
A method of separating large molecules (such as DNA fragments or proteins) from a mixture of similar molecules. An electric current is passed through a medium containing the mixture, and each kind of molecule travels through the medium at a different rate, depending on its electrical charge and size. Agarose and acrylamide gels are the media commonly used for electrophoresis of proteins and nucleic acids. (ORNL)

Electroporation
A process using high-voltage current to make cell membranes permeable to allow the introduction of new DNA; commonly used in recombinant DNA technology. *See also:* transfection. (ORNL)

Embryonic stem (ES) cells
An embryonic cell that can replicate indefinitely, transform into other types of cells, and serve as a continuous source of new cells. (ORNL)

Endonuclease
See: restriction enzyme. (ORNL)

Enzyme
A protein that acts as a catalyst, speeding the rate at which a biochemical reaction proceeds but not altering the direction or nature of the reaction. (ORNL)

Epistasis
One gene interferes with or prevents the expression of another gene located at a different locus. (ORNL)

Escherichia coli
Common bacterium that has been studied intensively by geneticists because of its small genome size, normal lack of pathogenicity, and ease of growth in the laboratory. (ORNL)

Eugenics
The study of improving a species by artificial selection; usually refers to the selective breeding of humans. (ORNL)

Eukaryote
Cell or organism with membrane-bound, structurally discrete nucleus and other well-developed subcellular compartments. Eukaryotes include all organisms except viruses, bacteria, and blue-green algae. *See also:* prokaryote, chromosome. (ORNL)

Evolutionarily conserved
See: conserved sequence. (ORNL)

Exogenous DNA
DNA originating outside an organism that has been introducted into the organism. (ORNL)

Exon
The protein-coding DNA sequence of a gene. *See also:* intron. (ORNL)

Exonuclease
An enzyme that cleaves nucleotides sequentially from free ends of a linear nucleic acid substrate. (ORNL)

Expressed gene
See: gene expression. (ORNL)

Expressed sequence tag (EST)
A short strand of DNA that is a part of a cDNA molecule and can act as identifier of a gene. Used in locating and mapping genes. *See also:* cDNA, sequence-tagged site. (ORNL)

F

FASTA
(a) The first widely used algorithm for database similarity searching. The program looks for optimal local alignments by scanning the sequence for small matches called "words." Initially, the scores of segments in which there are multiple word hits are calculated ("init1"). Later the scores of several segments may be summed to generate an "initn" score. An optimized alignment that includes gaps is shown in the output as "opt." The sensitivity and speed of the search are inversely related and controlled by the "k-tup" variable, which specifies the size of a word. (NCBI) (b) An output format for nucleic acid or protein sequences.

Filial generation (F1, F2)
Each generation of offspring in a breeding program, designated F1, F2, etc. (ORNL)

Filtering
Also known as masking. The process of hiding regions of (nucleic acid or amino acid) sequence having characteristics that frequently lead to spurious high scores. *See also:* SEG and DUST. (NCBI)

Fingerprinting
In genetics, the identification of multiple specific alleles on a person's DNA to produce a unique identifier for that person. *See also:* forensics. (ORNL)

Finished DNA sequence
High-quality, low-error, gap-free DNA sequence of the human genome. Achieving this ultimate 2003 Human Genome Project (HGP) goal requires additional sequencing to close gaps, reduce ambiguities, and allow for only a single error every 10,000 bases, the agreed-upon standard for HGP finished sequence. *See also:* sequencing, draft sequence. (ORNL)

Flow cytometry
Analysis of biological material by detection of the light-absorbing or fluorescing properties of cells or subcellular fractions (i.e., chromosomes) passing in a narrow stream through a

laser beam. An absorbance or fluorescence profile of the sample is produced. Automated sorting devices, used to fractionate samples, sort successive droplets of the analyzed stream into different fractions depending on the fluorescence emitted by each droplet. (ORNL)

Flow karyotyping
Use of flow cytometry to analyze and separate chromosomes according to their DNA content. (ORNL)

Fluorescence in situ hybridization (FISH)
A physical mapping approach that uses fluorescein tags to detect hybridization of probes with metaphase chromosomes and with the less-condensed somatic interphase chromatin. (ORNL)

Folded leaf
A layer of α helices wrapped around a single hydrophobic core but not with the simple geometry of a bundle. (SCOP)

Forensics
The use of DNA for identification. Some examples of DNA use are to establish paternity in child support cases, establish the presence of a suspect at a crime scene, and identify accident victims. (ORNL)

Fraternal twin
Siblings born at the same time as the result of fertilization of two ova by two sperm. They share the same genetic relationship to each other as any other siblings. *See also:* identical twin. (ORNL)

Full gene sequence
The complete order of bases in a gene. This order determines which protein a gene will produce. (ORNL)

Functional genomics
The study of genes, their resulting proteins, and the role played by the proteins in the body's biochemical processes. (ORNL)

G

Gamete
Mature male or female reproductive cell (sperm or ovum) with a haploid set of chromosomes (23 for humans). (ORNL)

Gap
(a) A space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment. (NCBI) (b) A position in an alignment that represents a deletion within one sequence relative to another. Gap penalties are requirements for alignment algorithms in order to reduce excessively gapped regions. Gaps in alignments represent insertions that usually occur in protruding loops or beta-bulges within protein structures. (SMART)

GC-rich area
Many DNA sequences carry long stretches of repeated G and C, which often indicates a gene-rich region. (ORNL)

Gel electrophoresis
See: electrophoresis. (ORNL)

Gene
The fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome that encodes a specific functional product (i.e., a protein or RNA molecule). *See also:* gene expression. (ORNL)

Gene amplification
Repeated copying of a piece of DNA; a characteristic of tumor cells. *See also:* gene, oncogene. (ORNL)

Gene chip technology
Development of cDNA microarrays from a large number of genes. Used to monitor and measure changes in gene expression for each gene represented on the chip. (ORNL)

Gene expression
The process by which a gene's coded information is converted into the structures present and operating in the cell. Expressed genes include those that are transcribed into mRNA and then translated into protein and those that are transcribed into RNA but not translated into protein (e.g., transfer and ribosomal RNAs). (ORNL)

Gene family
Group of closely related genes that make similar products. (ORNL)

Gene library
See: genomic library (ORNL)

Gene mapping
Determination of the relative positions of genes on a DNA molecule (chromosome or plasmid) and of the distance, in linkage units or physical units, between them. (ORNL)

Gene pool
All the variations of genes in a species. *See also:* allele, gene, polymorphism. (ORNL)

Gene prediction
Predictions of possible genes made by a computer program based on how well a stretch of DNA sequence matches known gene sequences. (ORNL)

Gene product
The biochemical material, either RNA or protein, resulting from expression of a gene. The amount of gene product is used to measure how active a gene is; abnormal amounts can be correlated with disease-causing alleles. (ORNL)

Gene testing
See: genetic testing, genetic screening. (ORNL)

Gene therapy
An experimental procedure aimed at replacing, manipulating, or supplementing nonfunctional or misfunctioning genes with healthy genes. *See also:* gene, inherit, somatic cell gene therapy, germ line gene therapy. (ORNL)

Genetic code
The sequence of nucleotides, coded in triplets (codons) along the mRNA, that determines the sequence of amino acids in protein synthesis. A gene's DNA sequence can be used to predict the mRNA sequence, and the genetic code can in turn be used to predict the amino acid sequence. (ORNL)

Genetic counseling
Provides patients and their families with education and information about genetic-related conditions and helps them make informed decisions. (ORNL)

Genetic discrimination
Prejudice against those who have or are likely to develop an inherited disorder. (ORNL)

Genetic engineering
Altering the genetic material of cells or organisms to enable them to make new substances or perform new functions. (ORNL)

Genetic engineering technology
See: recombinant DNA technology. (ORNL)

Genetic illness
Sickness, physical disability, or other disorder resulting from the inheritance of one or more deleterious alleles. (ORNL)

Genetic informatics
See: bioinformatics. (ORNL)

Genetic map
See: linkage map. (ORNL)

Genetic marker
A gene or other identifiable portion of DNA whose inheritance can be followed. *See also:* chromosome, DNA, gene, inherit. (ORNL)

Genetic material
See: genome. (ORNL)

Genetic mosaic
An organism in which different cells contain different genetic sequence. This can be the result of a mutation during development or fusion of embryos at an early developmental stage. (ORNL)

Genetic polymorphism
Difference in DNA sequence among individuals, groups, or populations (e.g., genes for blue eyes versus brown eyes). (ORNL)

Genetic predisposition
Susceptibility to a genetic disease. May or may not result in actual development of the disease. (ORNL)

Genetic screening
Testing a group of people to identify individuals at high risk of having or passing on a specific genetic disorder. (ORNL)

Genetic testing
Analyzing an individual's genetic material to determine predisposition to a particular health condition or to confirm a diagnosis of genetic disease. (ORNL)

Genetics
The study of inheritance patterns of specific traits. (ORNL)

Gene transfer
Incorporation of new DNA into an organism's cells, usually by a vector such as a modified virus. Used in gene therapy. *See also:* mutation, gene therapy, vector. (ORNL)

Genome
All the genetic material in the chromosomes of a particular organism; its size is generally given as its total number of base pairs. (ORNL)

Genome project
Research and technology development effort aimed at mapping and sequencing the genome of human beings and certain model organisms. *See also:* Human Genome Initiative. (ORNL)

Genomic library
A collection of clones made from a set of randomly generated overlapping DNA fragments that represent the entire genome of an organism. *See also:* library, arrayed library. (ORNL)

Genomics
The study of genes and their function. (ORNL)

Genomic sequence
See: DNA. (ORNL)

Genotype
The genetic constitution of an organism, as distinguished from its physical appearance (its phenotype). (ORNL)

Germ cell
Sperm and egg cells and their precursors. Germ cells are haploid and have only one set of chromosomes (23 in all), while all other cells have two copies (46 in all). (ORNL)

Germ line
The continuation of a set of genetic information from one generation to the next. *See also:* inherit. (ORNL)

Germ line gene therapy
An experimental process of inserting genes into germ cells or fertilized eggs to cause a genetic change that can be passed on to offspring. May be used to alleviate effects associated with a genetic disease. *See also:* genomics, somatic cell gene therapy. (ORNL)

Germ line genetic mutation
See: mutation. (ORNL)

Global alignment
The alignment of two nucleic acid or protein sequences over their entire length. (NCBI)

Greek key
A topology for a small number of β-sheet strands in which some interstrand connections go across the end of a barrel or, in a sandwich fold, between β sheets. (SCOP)

Guanine (G)
A nitrogenous base, one member of the base pair GC (guanine and cytosine) in DNA. *See also:* base pair, nucleotide. (ORNL)

H

H
The relative entropy of the target and background residue frequencies, H can be thought of as a measure of the average information (in bits) available per position that distinguishes an alignment from chance. At high values of H, short alignments can be distinguished by chance, whereas at lower H values, a longer alignment may be necessary. (NCBI)

Haploid
A single set of chromosomes (half the full set of genetic material) present in the egg and sperm cells of animals and in the egg and pollen cells of plants. Human beings have 23 chromosomes in their reproductive cells. *See also:* diploid. (ORNL)

Haplotype
A way of denoting the collective genotype of a number of closely linked loci on a chromosome. (ORNL)

Hemizygous
Having only one copy of a particular gene. For example, in humans, males are hemizygous for genes found on the Y chromosome. (ORNL)

Hereditary cancer
Cancer that occurs due to the inheritance of an altered gene within a family. *See also:* sporadic cancer. (ORNL)

Heterozygosity
The presence of different alleles at one or more loci on homologous chromosomes. (ORNL)

Heterozygote
See: heterozygosity. (ORNL)

Highly conserved sequence
DNA sequence that is very similar across several different types of organisms. *See also:* gene, mutation. (ORNL)

High-throughput sequencing
A fast method of determining the order of bases in DNA. *See also:* sequencing. (ORNL)

Homeobox
A short stretch of nucleotides whose base sequence is virtually identical in all the genes that contain it. Homeoboxes have been found in many organisms from fruit flies to human beings. In the fruit fly, a homeobox appears to determine when particular groups of genes are expressed during development. (ORNL)

Homolog
A member of a chromosome pair in diploid organisms or a gene that has the same origin and functions in two or more species. (ORNL)

Homologous chromosome
Chromosome containing the same linear gene sequences as another, each derived from one parent. (ORNL)

Homologous recombination
Swapping of DNA fragments between paired chromosomes. (ORNL)

Homology
(a) Similarity attributed to descent from a common ancestor. (NCBI) (b) Similarity in DNA or protein sequences between individuals of the same species or among different species. (ORNL) (c) Evolutionary descent from a common ancestor due to gene duplication. (SMART)

Homozygote
An organism that has two identical alleles of a gene. *See also:* heterozygote. (ORNL)

Homozygous
See: homozygote. (ORNL)

HSP
High-scoring segment pair. Local alignments with no gaps that achieve one of the top alignment scores in a given search. (NCBI)

Human gene therapy
See: gene therapy. (ORNL)

Human Genome Initiative
Collective name for several projects begun in 1986 by the U.S. Department of Energy (DOE) to create an ordered set of DNA segments from known chromosomal locations, develop new computational methods for analyzing genetic map and DNA sequence data, and develop new techniques and instruments for derecting and analyzing DNA. This DOE initiative is now known as the Human Genome Program. The joint national effort, led by the DOE and National Institutes of Health, is known as the Human Genome Project. (ORNL)

Human Genome Project (HGP)
Formerly titled Human Genome Initiative. *See also:* Human Genome Initiative. (ORNL)

Hybrid
The offspring of genetically different parents. *See also:* heterozygote. (ORNL)

Hybridization
The process of joining two complementary strands of DNA or one each of DNA and RNA to form a double-stranded molecule. (ORNL)

I

Identical twin
Twins produced by the division of a single zygote; both have identical genotypes. *See also:* fraternal twin. (ORNL)

Identity
The extent to which two (nucleotide or amino acid) sequences are invariant. (NCBI)

Immunotherapy
Using the immune system to treat disease, for example, in the development of vaccines. May also refer to the therapy of diseases caused by the immune system. *See also:* cancer. (ORNL)

Imprinting
A phenomenon in which the disease phenotype depends on which parent passed on the disease gene. For instance, both Prader-Willi and Angelman syndromes are inherited when the same part of chromosome 15 is missing. When the father's complement of 15 is missing, the child has Prader-Willi, but when the mother's complement of 15 is missing, the child has Angelman syndrome. (ORNL)

Independent assortment
During meiosis each of the two copies of a gene is distributed to the germ cells independently of the distribution of other genes. *See also:* linkage. (ORNL)

Informatics
See: bioinformatics. (ORNL)

Informed consent
An individual willingly agrees to participate in an activity after first being advised of the risks and benefits. *See also:* privacy. (ORNL)

Inherit
In genetics, to receive genetic material from parents through biological processes. (ORNL)

Inherited
See: inherit. (ORNL)

Insertion
A chromosome abnormality in which a piece of DNA is incorporated into a gene and thereby disrupts the gene's normal function. *See also:* chromosome, DNA, gene, mutation. (ORNL)

Insertional mutation
See: insertion. (ORNL)

In situ hybridization
Use of a DNA or RNA probe to detect the presence of the complementary DNA sequence in cloned bacterial or cultured eukaryotic cells. (ORNL)

Intellectual property rights
Patents, copyrights, and trademarks. *See also:* patent. (ORNL)

Interference
One crossover event inhibits the chances of another crossover event. Also known as positive interference. Negative interference increases the chance of a second crossover. *See also:* crossing over. (ORNL)

Interphase
The period in the cell cycle when DNA is replicated in the nucleus; followed by mitosis. (ORNL)

Intracellular domains
Domain families that are most prevalent in proteins within the cytoplasm. (SMART)

Intron
DNA sequence that interrupts the protein-coding sequence of a gene; an intron is transcribed into RNA but is cut out of the message before it is translated into protein. *See also:* exon. (ORNL)

In vitro
Studies performed outside a living organism, such as in a laboratory. (ORNL)

In vivo
Studies carried out in living organisms. (ORNL)

Isoenzyme
An enzyme performing the same function as another enzyme but having a different set of amino acids. The two enzymes may function at different speeds. (ORNL)

J

Jelly roll
A variant of Greek-key topology with both ends of a sandwich or a barrel fold being crossed by two interstrand connections. *See also:* Greek key. (SCOP)

Junk DNA
Stretches of DNA that do not code for genes; most of the genome consists of so-called junk DNA which may have regulatory and other functions. Also called noncoding DNA. (ORNL)

K

K
A statistical parameter used in calculating BLAST scores that can be thought of as a natural scale for search space size. The value K is used in converting a raw score (S) to a bit score (S'). (NCBI)

Karyotype
A photomicrograph of an individual's chromosomes arranged in a standard format showing the number, size, and shape of each chromosome type; used in low-resolution physical mapping to correlate gross chromosomal abnormalities with the characteristics of specific diseases. (ORNL)

Kilobase (kb)
Unit of length for DNA fragments equal to 1000 nucleotides. (ORNL)

Knock-out
Deactivation of specific genes; used in laboratory organisms to study gene function. *See also:* gene, locus, model organisms. (ORNL)

L

Lambda
A statistical parameter used in calculating BLAST scores that can be thought of as a natural scale for a scoring system. The value lambda is used in converting a raw score (S) to a bit score (S'). (NCBI)

Library
An unordered collection of clones (i.e., cloned DNA from a particular organism) whose relationship to each other can be established by physical mapping. *See also:* genomic library, arrayed library. (ORNL)

Linkage
The proximity of two or more markers (e.g., genes, restriction fragment length polymorphism markers) on a chromosome; the closer the markers, the lower the probability that they will be separated during DNA repair or replication processes (binary fission in prokaryotes, mitosis or meiosis in eukaryotes), and hence the greater the probability that they will be inherited together. (ORNL)

Linkage disequilibrium
Where alleles occur together more often than can be accounted for by chance. Indicates that the two alleles are physically close on the DNA strand. *See also:* Mendelian inheritance. (ORNL)

Linkage map
A map of the relative positions of genetic loci on a chromosome, determined on the basis of how often the loci are inherited together. Distance is measured in centimorgans (cM). (ORNL)

Local alignment
The alignment of some portion of two nucleic acid or protein sequences. (NCBI)

Localization
Numbers of domains that are thought from SwissProt annotations to be present in different cellular compartments

(cytoplasm, extracellular space, nucleus, and membrane associated) are shown in annotation pages. (SMART)

Localize
Determination of the original position (locus) of a gene or other marker on a chromosome. (ORNL)

Locus (plural loci)
The position on a chromosome of a gene or other chromosome marker; also, the DNA at that position. The use of locus is sometimes restricted to mean expressed DNA regions. *See also:* gene expression. (ORNL)

Long-range restriction mapping
Restriction enzymes are proteins that cut DNA at precise locations. Restriction maps depict the chromosomal positions of restriction enzyme cutting sites. These are used as biochemical "signposts," or markers of specific areas along the chromosomes. The map will detail the positions where the DNA molecule is cut by particular restriction enzymes. (ORNL)

Low-complexity region (LCR)
Regions of biased composition including homopolymeric runs, short-period repeats, and more subtle overrepresentation of one or a few residues. The SEG program is used to mask or filter LCRs in amino acid queries. The DUST program is used to mask or filter LCRs in nucleic acid queries. (NCBI)

M

Macrorestriction map
Map depicting the order of and distance between sites at which restriction enzymes cleave chromosomes. (ORNL)

Mapping
See: gene mapping, linkage map, physical map. (ORNL)

Mapping population
The group of related organisms used in constructing a genetic map. (ORNL)

Marker
See: genetic marker. (ORNL)

Masking
Also known as filtering. The removal of repeated or low-complexity regions from a sequence in order to improve the sensitivity of sequence similarity searches performed with that sequence. (NCBI)

Mass spectrometer
An instrument used to identify chemicals in a substance by their mass and charge. (ORNL)

Meander
A simple topology of a β sheet where any two consecutive strands are adjacent and antiparallel. (SCOP)

Megabase (Mb)
Unit of length for DNA fragments equal to 1 million nucleotides and roughly equal to 1 cM. *See also:* centimorgan. (ORNL)

Meiosis
The process of two consecutive cell divisions in the diploid progenitors of sex cells. Meiosis results in four rather than two daughter cells, each with a haploid set of chromosomes. *See also:* mitosis. (ORNL)

Mendelian inheritance
One method in which genetic traits are passed from parents to offspring. Named for Gregor Mendel, who first studied and recognized the existence of genes and this method of inheritance. *See also:* autosomal dominant, recessive gene, sex linked. (ORNL)

Messenger RNA (mRNA)
RNA that serves as a template for protein synthesis. *See also:* genetic code. (ORNL)

Metaphase
A stage in mitosis or meiosis during which the chromosomes are aligned along the equatorial plane of the cell. (ORNL)

Microarray
Sets of miniaturized chemical reaction areas that may also be used to test DNA fragments, antibodies, or proteins. (ORNL)

Microbial genetics
The study of genes and gene function in bacteria, archaea, and other microorganisms. Often used in research in the fields of bioremediation, alternative energy, and disease prevention. *See also:* model organisms, biotechnology, bioremediation. (ORNL)

Microinjection
A technique for introducing a solution of DNA into a cell using a fine microcapillary pipet. (ORNL)

Mitochondrial DNA
The genetic material found in mitochondria, the organelles that generate energy for the cell. Not inherited in the same fashion as nucleic DNA. *See also:* cell, DNA, genome, nucleus. (ORNL)

Mitosis
The process of nuclear division in cells that produces daughter cells that are genetically identical to each other and to the parent cell. *See also:* meiosis. (ORNL)

Modeling
The use of statistical analysis, computer analysis, or model organisms to predict outcomes of research. (ORNL)

Model organisms
A laboratory animal or other organism useful for research. (ORNL)

Molecular biology
The study of the structure, function, and makeup of biologically important molecules. (ORNL)

Molecular farming
The development of transgenic animals to produce human proteins for medical use. (ORNL)

Molecular genetics
The study of macromolecules important in biological inheritance. (ORNL)

Molecular medicine
The treatment of injury or disease at the molecular level. Examples include the use of DNA-based diagnostic tests or medicine derived from DNA sequence information. (ORNL)

Monogenic disorder
A disorder caused by mutation of a single gene. *See also:* mutation, polygenic disorder. (ORNL)

Monogenic inheritance
See: monogenic disorder. (ORNL)

Monosomy
Possessing only one copy of a particular chromosome instead of the normal two copies. *See also:* cell, chromosome, gene expression, trisomy. (ORNL)

Morbid map
A diagram showing the chromosomal location of genes associated with disease. (ORNL)

Motif
(a) A short conserved region in a protein sequence. Motifs are frequently highly conserved parts of domains. (NCBI) (b) Sequence motifs are short conserved regions of polypeptides. Sets of sequence motifs need not necessarily represent homologs. (SMART)

Mouse model
See: model organisms. (ORNL)

Multifactorial or multigenic disorder
See: polygenic disorder. (ORNL)

Multiple sequence alignment
An alignment of three or more sequences with gaps inserted in the sequences such that residues with common structural positions and/or ancestral residues are aligned in the same column. ClustalW is one of the most widely used multiple sequence alignment programs. (NCBI)

Multiplexing
A laboratory approach that performs multiple sets of reactions in parallel (simultaneously); greatly increasing speed and throughput. (ORNL)

Murine
Organism in the genus *Mus.* A rat or mouse. (ORNL)

Mutagen
An agent that causes a permanent genetic change in a cell. Does not include changes occurring during normal genetic recombination. (ORNL)

Mutagenicity
The capacity of a chemical or physical agent to cause permanent genetic alterations. *See also:* somatic cell genetic mutation. (ORNL)

Mutation
Any heritable change in DNA sequence. *See also:* polymorphism. (ORNL)

N

Nitrogenous base
A nitrogen-containing molecule having the chemical properties of a base. DNA contains the nitrogenous bases adenine (A), guanine (G), cytosine (C), and thymine (T). *See also:* DNA. (ORNL)

Northern blot
A gel-based laboratory procedure that locates mRNA sequences on a gel that are complementary to a piece of DNA used as a probe. *See also:* DNA, library. (ORNL)

Nuclear transfer
A laboratory procedure in which a cell's nucleus is removed and placed into an oocyte with its own nucleus removed so the genetic information from the donor nucleus controls the resulting cell. Such cells can be induced to form embryos. This process was used to create the cloned sheep Dolly. *See also:* cloning. (ORNL)

Nucleic acid
A large molecule composed of nucleotide subunits. *See also:* DNA. (ORNL)

Nucleolar organizing region
A part of the chromosome containing rRNA genes. (ORNL)

Nucleotide
A subunit of DNA or RNA consisting of a nitrogenous base (adenine, guanine, thymine, or cytosine in DNA; adenine, guanine, uracil, or cytosine in RNA), a phosphate molecule, and a sugar molecule (deoxyribose in DNA and ribose in RNA). Thousands of nucleotides are linked to form a DNA or RNA molecule. *See also:* DNA, base pair, RNA. (ORNL)

Nucleus
The cellular organelle in eukaryotes that contains most of the genetic material. (ORNL)

O

Oligo
See: oligonucleotide. (ORNL)

Oligogenic
A phenotypic trait produced by two or more genes working together. *See also:* polygenic disorder. (ORNL)

Oligonucleotide
A molecule usually composed of 25 or fewer nucleotides; used as a DNA synthesis primer. *See also:* nucleotide. (ORNL)

Oncogene
A gene, one or more forms of which are associated with cancer. Many oncogenes are involved, directly or indirectly, in controlling the rate of cell growth. (ORNL)

Open reading frame (ORF)
The sequence of DNA or RNA located between the start-code sequence (initiation codon) and the stop-code sequence (termination codon). (ORNL)

Operon
A set of genes transcribed under the control of an operator gene. (ORNL)

Optimal alignment
An alignment of two sequences with the highest possible score. (NCBI)

ORF
See: open reading frame. (SMART)

Orthologous
Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function. (NCBI)

Overlapping clones
See: genomic library. (ORNL)

P

***P* value**
The probability of an alignment occurring with the score in question or better. The *P* value is calculated by relating the observed alignment score, *S*, to the expected distribution of high-scoring segment pair scores from comparisons of random sequences of the same length and composition as the query to the database. The most highly significant *P* values will be those close to zero. The *P* and *E* values are different ways of representing the significance of the alignment. (NCBI)

P1-derived artificial chromosome (PAC)
One type of vector used to clone DNA fragments (insert size 100–300 kb; average 150 kb) in *Escherichia coli* cells. Based on bacteriophage (a virus) P1 genome. *See also:* cloning vector. (ORNL)

PAM
Point accepted mutation. A unit used to quantify the amount of evolutionary change in a protein sequence. The amount of evolution which will change, on average, 1% of amino acids in a protein sequence is 1.0 PAM units. A PAM(x) substitution matrix is a look-up table in which scores for each amino acid substitution have been calculated based on the frequency of that substitution in closely related proteins that have experienced a certain amount (x) of evolutionary divergence. (NCBI)

Paralogous
Homologous sequences within a single species that arose by gene duplication. (NCBI)

Partly open barrel
Has the edge strands not properly hydrogen bonded because one of the strands is in two parts connected with a linker of more than that one residue. These edge strands can be treated as a single but interrupted strand, allowing classification with the effective strand and shear numbers, n^* and S^*. In the few open barrels the β sheets are connected by only a few *side-chain* hydrogen bonds between the edge strands. (SCOP)

Patent
In genetics, conferring the right or title to genes, gene variations, or identifiable portions of sequenced genetic material to an individual or organization. *See also:* gene. (ORNL)

Pedigree
A family tree diagram that shows how a particular genetic trait or disease has been inherited. *See also:* inherit. (ORNL)

Penetrance
The probability of a gene or genetic trait being expressed. "Complete" penetrance means the gene or genes for a trait are expressed in the whole population that has the genes. "Incomplete" penetrance means the genetic trait is expressed in only part of the population. The percent penetrance also may change with the age range of the population. (ORNL)

Peptide
Two or more amino acids joined by a bond called a "peptide bond." *See also:* polypeptide. (ORNL)

Phage
A virus for which the natural host is a bacterial cell. (ORNL)

Pharmacogenomics
The study of the interaction of an individual's genetic makeup and response to a drug. (ORNL)

Phenocopy
A trait not caused by inheritance of a gene but that appears to be identical to a genetic trait. (ORNL)

Phenotype
The physical characteristics of an organism or the presence of a disease that may or may not be genetic. *See also:* genotype. (ORNL)

Physical map
A map of the locations of identifiable landmarks on DNA (e.g., restriction enzyme cutting sites, genes), regardless of inheritance. Distance is measured in base pairs. For the human genome, the lowest resolution physical map is the banding patterns on the 24 different chromosomes; the highest resolution map is the complete nucleotide sequence of the chromosomes. (ORNL)

Plasmid
Autonomously replicating extrachromosomal circular DNA molecules, distinct from the normal bacterial genome and nonessential for cell survival under nonselective conditions. Some plasmids are capable of integrating into the host genome. A number of artificially constructed plasmids are used as cloning vectors. (ORNL)

Pleiotropy
One gene that causes many different physical traits such as multiple disease symptoms. (ORNL)

Pluripotency
The potential of a cell to develop into more than one type of mature cell, depending on environment. (ORNL)

Polygenic disorder
Genetic disorder resulting from the combined action of alleles of more than one gene (e.g., heart disease, diabetes, and some cancers). Although such disorders are inherited, they depend on the simultaneous presence of several alleles; thus the hereditary patterns usually are more complex than those of single-gene disorders. *See also:* single-gene disorder. (ORNL)

Polymerase chain reaction (PCR)
A method for amplifying a DNA base sequence using a heat-stable polymerase and two 20-base primers, one complementary to the (+) strand at one end of the sequence to be amplified and one complementary to the (−) strand at the other end. Because the newly synthesized DNA strands can subsequently serve as additional templates for the same primer sequences, successive rounds of primer annealing, strand elongation, and dissociation produce rapid and highly specific amplification of the desired

sequence. PCR also can be used to detect the existence of the defined sequence in a DNA sample. (ORNL)

Polymerase, DNA or RNA
Enzyme that catalyzes the synthesis of nucleic acids on preexisting nucleic acid templates, assembling RNA from ribonucleotides or DNA from deoxyribonucleotides. (ORNL)

Polymorphism
Difference in DNA sequence among individuals that may underlie differences in health. Genetic variations occurring in more than 1% of a population would be considered useful polymorphisms for genetic linkage analysis. *See also:* mutation. (ORNL)

Polypeptide
A protein or part of a protein made of a chain of amino acids joined by a peptide bond. (ORNL)

Population genetics
The study of variation in genes among a group of individuals. (ORNL)

Positional cloning
A technique used to identify genes, usually those that are associated with diseases, based on their location on a chromosome. (ORNL)

Primer
Short preexisting polynucleotide chain to which new deoxyribonucleotides can be added by DNA polymerase. (ORNL)

Privacy
In genetics, the right of people to restrict access to their genetic information. (ORNL)

Probe
Single-stranded DNA or RNA molecules of specific base sequence, labeled either radioactively or immunologically, that are used to detect the complementary base sequence by hybridization. (ORNL)

Profile
(a) A table that lists the frequencies of each amino acid in each position of protein sequence. Frequencies are calculated from multiple alignments of sequences containing a domain of interest. *See also:* PSSM. (NCBI) (b) A table of position-specific scores and gap penalties, representing an homologous family, that may be used to search sequence databases. In ClustalW-derived profiles those sequences that are more distantly related are assigned higher weights. (SMART)

Prokaryote
Cell or organism lacking a membrane-bound, structurally discrete nucleus and orher subcellular compartments. Bacteria are examples of prokaryotes. *See also:* chromosome, eukaryote. (ORNL)

Promoter
A DNA site to which RNA polymerase will bind and initiate transcription. (ORNL)

Pronucleus
The nucleus of a sperm or egg prior to fertilization. *See also:* nucleus, transgenic. (ORNL)

Protein
A large molecule composed of one or more chains of amino acids in a specific order; the order is determined by the base sequence of nucleotides in the gene that codes for the protein. Proteins are required for the structure, function, and regulation of the body's cells, tissues, and organs; each protein has unique functions. Examples are hormones, enzymes, and antibodies. (ORNL)

Proteome
Proteins expressed by a cell or organ at a particular time and under specific conditions. (ORNL)

Proteomics
Systematic analysis of protein expression of normal and diseased tissues that involves the separation, identification, and characterization of all of the proteins in an organism. (NCBI)

Pseudogene
A sequence of DNA similar to a gene but nonfunctional; probably the remnant of a once functional gene that accumulated mutations. (ORNL)

PSI-BLAST
Position-Specific Iterative BLAST. An iterative search using the BLAST algorithm. A profile is built after the initial search, which is then used in subsequent searches. The process may be repeated, if desired, with new sequences found in each cycle used to refine the profile. (NCBI)

PSSM
Position-specific scoring matrix. The PSSM gives the log-odds score for finding a particular matching amino acid in a target sequence. *See also:* profile. (NCBI)

Purine
A nitrogen-containing, double-ring, basic compound that occurs in nucleic acids. The purines in DNA and RNA are adenine and guanine. *See also:* base pair. (ORNL)

Pyrimidine
A nitrogen-containing, single-ring, basic compound that occurs in nucleic acids. The pyrimidines in DNA are cytosine and thymine; in RNA, cytosine and uracil. *See also:* base pair. (ORNL)

Q

Query
The input sequence (or other type of search term) with which all of the entries in a database are to be compared. (NCBI)

R

Radiation hybrid
A hybrid cell containing small fragments of irradiated human chromosomes. Maps of irradiation sites on chromosomes for the human, rat, mouse, and other genomes provide important markers, allowing the construction of very precise sequence-tagged site maps indispensable to studying multifactorial diseases. *See also:* sequence-tagged site. (ORNL)

Rare-cutter enzyme
See: restriction enzyme cutting site. (ORNL)

Raw score
The score of an alignment, *S*, calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table. Gap scores are typically calculated as the sum of *G*, the gap opening penalty, and *L*, the gap extension penalty. For a gap of length *n*, the gap cost would be $G + Ln$. The choice of gap costs *G* and *L* is empirical, but it is customary to choose a high value for *G* (10–15) and a low value for *L* (1–2). *See also:* PAM, BLOSUM. (NCBI)

Recessive gene
A gene which will be expressed only if there are two identical copies or, for a male, if one copy is present on the X chromosome. (ORNL)

Reciprocal translocation
When a pair of chromosomes exchange exactly the same length and area of DNA. Results in a shuffling of genes. (ORNL)

Recombinant clone
Clone containing recombinant DNA molecules. *See also:* recombinant DNA technology. (ORNL)

Recombinant DNA molecules
A combination of DNA molecules of different origin that are joined using recombinant DNA technologies. (ORNL)

Recombinant DNA technology
Procedure used to join together DNA segments in a cell-free system (an environment outside a cell or organism). Under appropriate conditions, a recombinant DNA molecule can enter a cell and replicate there, either autonomously or after it has become integrated into a cellular chromosome. (ORNL)

Recombination
The process by which progeny derives a combination of genes different from that of either parent. In higher organisms, this can occur by crossing over. *See also:* crossing over, mutation. (ORNL)

Regulatory region or sequence
A DNA base sequence that controls gene expression. (ORNL)

Repetitive DNA
Sequences of varying lengths that occur in multiple copies in the genome; it represents much of the human genome. (ORNL)

Reporter gene
See: marker. (ORNL)

Resolution
Degree of molecular detail on a physical map of DNA, ranging from low to high. (ORNL)

Restriction enzyme cutting site
A specific nucleotide sequence of DNA at which a particular restriction enzyme cuts the DNA. Some sites occur frequently in DNA (e.g., every several hundred base pairs); others much less frequently (rare cutter; e.g., every 10,000 bp). (ORNL)

Restriction enzyme, endonuclease
A protein that recognizes specific, short nucleotide sequences and cuts DNA at those sites. Bacteria contain over 400 such enzymes that recognize and cut more than 100 different DNA sequences. *See also:* restriction enzyme cutting site. (ORNL)

Restriction fragment length polymorphism (RFLP)
Variation between individuals in DNA fragment sizes cut by specific restriction enzymes; polymorphic sequences that result in RFLPs are used as markers on both physical maps and genetic linkage maps. RFLPs usually are caused by mutation at a cutting site. *See also:* marker, polymorphism. (ORNL)

Retroviral infection
The presence of retroviral vectors, such as some viruses, which use their recombinant DNA to insert their genetic material into the chromosomes of the host's cells. The virus is then propagated by the host cell. (ORNL)

Reverse transcriptase
An enzyme used by retroviruses to form a complementary DNA sequence (cDNA) from their RNA. The resulting DNA is then inserted into the chromosome of the host cell. (ORNL)

Ribonucleotide
See: nucleotide. (ORNL)

Ribose
The five-carbon sugar that serves as a component of RNA. *See also:* ribonucleic acid, deoxyribose. (ORNL)

Ribosomal RNA (rRNA)
A class of RNA found in the ribosomes of cells. (ORNL)

Ribosomes
Small cellular components composed of specialized ribosomal RNA and protein; site of protein synthesis. *See also:* RNA. (ORNL)

Risk communication
In genetics, a process in which a genetic counselor or other medical professional interprets genetic test results and advises patients of the consequences for them and their offspring. (ORNL)

RNA (ribonucleic acid)
A chemical found in the nucleus and cytoplasm of cells; it plays an important role in protein synthesis and other chemical activities of the cell. The structure of RNA is similar to that of DNA. There are several classes of RNA molecules, including messenger RNA, transfer RNA, ribosomal RNA, and other small RNAs, each serving a different purpose. (ORNL)

S

Sanger sequencing
A widely used method of determining the order of bases in DNA. *See also:* sequencing, shotgun sequencing. (ORNL)

Satellite
A chromosomal segment that branches off from the rest of the chromosome but is still connected by a thin filament or stalk. (ORNL)

Scaffold
In genomic mapping, a series of contigs that are in the right order but not necessarily connected in one continuous stretch of sequence. (ORNL)

Seed alignment
Alignment that contains only one of each pair of homologs that are represented in a ClustalW-derived phylogenetic tree linked by a branch of length less than a distance of 0.2. (SMART)

SEG
A program for filtering low-complexity regions in amino acid sequences. Residues that have been masked are represented as "X" in an alignment. SEG filtering is performed by default in the blastp subroutine of BLAST 2.0. (NCBI)

Segregation
The normal biological process whereby the two pieces of a chromosome pair are separated during meiosis and randomly distributed to the germ cells. (ORNL)

Sequence
See: base sequence. (ORNL)

Sequence assembly
A process whereby the order of multiple sequenced DNA fragments is determined. (ORNL)

Sequence-tagged site (STS)
Short (200–500-bp) DNA sequence that has a single occurrence in the human genome and whose location and base sequence are known. Detectable by polymerase chain reaction, STSs are useful for localizing and orienting the mapping and sequence data reported from many different laboratories and serve as landmarks on the developing physical map of the human genome. Expressed sequence tags (ESTs) are STSs derived from cDNAs. (ORNL)

Sequencing
Determination of the order of nucleotides (base sequences) in a DNA or RNA molecule or the order of amino acids in a protein. (ORNL)

Sequencing technology
The instrumentation and procedures used to determine the order of nucleotides in DNA. (ORNL)

Sex chromosome
The X or Y chromosome in human beings that determines the sex of an individual. Females have two X chromosomes in diploid cells; males have an X and a Y chromosome. The sex chromosomes comprise the 23rd chromosome pair in a karyotype. *See also:* autosome. (ORNL)

Sex linked
Traits or diseases associated with the X or Y chromosome; generally seen in males. *See also:* gene, mutation, sex chromosome. (ORNL)

Shotgun method
Sequencing method that involves randomly sequenced cloned pieces of the genome, with no foreknowledge of where the piece originally came from. This can be contrasted with "directed" strategies, in which pieces of DNA from known chromosomal locations are sequenced. Because there are advantages to both strategies, researchers use both random (or shotgun) and directed strategies in combination to sequence the human genome. *See also:* library, genomic library. (ORNL)

Similarity
The extent to which nucleotide or protein sequences are related. The extent of similarity between two sequences can be based on percent sequence identity and/or conservation. In BLAST similarity refers to a positive matrix score. (NCBI)

Single-gene disorder
Hereditary disorder caused by a mutant allele of a single gene (e.g., Duchenne muscular dystrophy, retinoblastoma, sickle cell disease). *See also:* polygenic disorders. (ORNL)

Single-nucleotide polymorphism (SNP)
DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. *See also:* mutation, polymorphism, single-gene disorder. (ORNL)

Somatic cell
Any cell in the body except gametes and their precursors. *See also:* gamete. (ORNL)

Somatic cell gene therapy
Incorporating new genetic material into cells for therapeutic purposes. The new genetic material cannot be passed to offspring. *See also:* gene therapy. (ORNL)

Somatic cell genetic mutation
A change in the genetic structure that is neither inherited nor passed to offspring. Also called acquired mutations. *See also:* germ line genetic mutation. (ORNL)

Southern blotting
Transfer by absorption of DNA fragments separated in electrophoretic gels to membrane filters for detection of specific base sequences by radiolabeled complementary probes. (ORNL)

Spectral karyotype (SKY)
A graphic of all an organism's chromosomes, each labeled with a different color. Useful for identifying chromosomal abnormalities. *See also:* chromosome. (ORNL)

Splice site
Location in the DNA sequence where RNA removes the noncoding areas to form a continuous gene transcript for translation into a protein. (ORNL)

Sporadic cancer
Cancer that occurs randomly and is not inherited from parents. Caused by DNA changes in one cell that grows and divides, spreading throughout the body. *See also:* hereditary cancer. (ORNL)

Stem cell
Undifferentiated, primitive cells in the bone marrow that have the ability both to multiply and to differentiate into specific blood cells. (ORNL)

Structural genomics
The effort to determine the three-dimensional structures of large numbers of proteins using both experimental techniques and computer simulation. (ORNL)

Substitution
(a) The presence of a nonidentical amino acid at a given position in an alignment. If the aligned residues have similar physicochemical properties, the substitution is said to be "conservative." (NCBI) (b) In genetics, a type of mutation due to replacement of

one nucleotide in a DNA sequence by another nucleotide or replacement of one amino acid in a protein by another amino acid. *See also:* mutation. (ORNL)

Substitution matrix
A substitution matrix containing values proportional to the probability that amino acid *i* mutates into amino acid *j* for all pairs of amino acids. Such matrices are constructed by assembling a large and diverse sample of verified pairwise alignments of amino acids. If the sample is large enough to be statistically significant, the resulting matrices should reflect the true probabilities of mutations occurring through a period of evolution. (NCBI)

Suppressor gene
A gene that can suppress the action of another gene. (ORNL)

Syndrome
The group or recognizable pattern of symptoms or abnormalities that indicate a particular trait or disease. (ORNL)

Syngeneic
Genetically identical members of the same species. (ORNL)

Synteny
Genes occurring in the same order on chromosomes of different species. *See also:* linkage, conserved sequence. (ORNL)

T

Tandem repeat sequences
Multiple copies of the same base sequence on a chromosome; used as markers in physical mapping. *See also:* physical map. (ORNL)

Targeted mutagenesis
Deliberate change in the genetic structure directed at a specific site on the chromosome. Used in research to determine the targeted region's function. *See also:* mutation, polymorphism. (ORNL)

Technology transfer
The process of transferring scientific findings from research laboratories to the commercial sector. (ORNL)

Telomerase
The enzyme that directs the replication of telomeres. (ORNL)

Telomere
The end of a chromosome. This specialized structure is involved in the replication and stability of linear DNA molecules. *See also:* DNA replication. (ORNL)

Teratogenic
Substances such as chemicals or radiation that causes abnormal development of an embryo. *See also:* mutatgen. (ORNL)

Thymine (T)
A nitrogenous base, one member of the base pair AT (adenine-thymine). *See also:* base pair, nucleotide. (ORNL)

Toxicogenomics
The study of how genomes respond to environmental stressors or toxicants. Combines genomewide mRNA expression profiling with protein expression patterns using bioinformatics to understand the role of gene–environment interactions in disease and dysfunction. (ORNL)

Transcription
The synthesis of an RNA copy from a sequence of DNA (a gene); the first step in gene expression. *See also:* translation. (ORNL)

Transcription factor
A protein that binds to regulatory regions and helps control gene expression. (ORNL)

Transcriptome
The full complement of activated genes, mRNAs, or transcripts in a particular tissue at a particular time. (ORNL)

Transfection
The introduction of foreign DNA into a host cell. *See also:* cloning vector, gene therapy. (ORNL)

Transfer RNA (tRNA)
A class of RNA having structures with triplet nucleotide sequences that are complementary to the triplet nucleotide coding sequences of mRNA. The role of tRNAs in protein synthesis is to bond with amino acids and transfer them to the ribosomes, where proteins are assembled according to the genetic code carried by mRNA. (ORNL)

Transformation
A process by which the genetic material carried by an individual cell is altered by incorporation of exogenous DNA into its genome. (ORNL)

Transgenic
An experimentally produced organism in which DNA has been artificially introduced and incorporated into the organism's germ line. *See also:* cell, DNA, gene, nucleus, germ line. (ORNL)

Translation
The process in which the genetic code carried by mRNA directs the synthesis of proteins from amino acids. *See also:* transcription. (ORNL)

Translocation
A mutation in which a large segment of one chromosome breaks off and attaches to another chromosome. *See also:* mutation. (ORNL)

Transposable element
A class of DNA sequences that can move from one chromosomal site to another. (ORNL)

Trisomy
Possessing three copies of a particular chromosome instead of the normal two copies. *See also:* cell, gene, gene expression, chromosome. (ORNL)

U

Unitary matrix
Also known as identity matrix. A scoring system in which only identical characters receive a positive score. (NCBI)

Up and down
The simplest topology for a helical bundle or folded leaf, in which consecutive helices are adjacent and antiparallel; it is approximately equivalent to the meander topology of a β sheet. (SCOP)

Uracil
A nitrogenous base normally found in RNA but not DNA; it is capable of forming a base pair with adenine. *See also:* base pair, nucleotide. (ORNL)

V

Vector
See: cloning vector. (ORNL)

Virus
A noncellular biological entity that can reproduce only within a host cell. Viruses consist of nucleic acid covered by protein; some animal viruses are also surrounded by membrane. Inside the infected cell, the virus uses the synthetic capability of the host to produce progeny virus. *See also:* cloning vector. (ORNL)

W

Western blot
A technique used to identify and locate proteins based on their ability to bind to specific antibodies. *See also:* DNA, Northern blot, protein, RNA, Southern blotting. (ORNL)

Wild type
The form of an organism that occurs most frequently in nature. (ORNL)

Working draft DNA sequence
See: Draft DNA sequence. (ORNL)

X

X chromosome
One of the two sex chromosomes, X and Y. *See also:* Y chromosome, sex chromosome. (ORNL)

Xenograft
Tissue or organs from an individual of one species transplanted into or grafted onto an organism of another species, genus, or family. A common example is the use of pig heart valves in humans. (ORNL)

Y

Y chromosome
One of the two sex chromosomes, X and Y. *See also:* X chromosome, sex chromosome. (ORNL)

Yeast artificial chromosome (YAC)
Constructed from yeast DNA, it is a vector used to clone large DNA fragments. *See also:* cloning vector, cosmid. (ORNL)

Z

Zinc-finger protein
A secondary feature of some proteins containing a zinc atom; a DNA-binding protein. (ORNL)

TABLE 1 Glossaries Available on Internet

Source	URL
RNA glossary from Cambridge Healthtech Institute	http://www.genomicglossaries.com/content/RNA.asp
Glossary of Biotechnology Terms	http://smart.embl-heidelberg.de/help/smart_glossary.shtml
Glossary of structural genomics	http://scop.mrc-lmb.cam.ac.uk/scop/gloss.html
Genome Glossary (Oak Ridge National Laboratory)	http://www.ornl.gov/TechResources/Human_Genome/glossary/
The National Human Genome Research Institute (NHGRI), Talking Glossary of Genetic Terms	http://www.genome.gov/glossary.cfm
BLAST glossary from NCBI	http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html
Genomics Glossary from Cambridge Healthtech Institute	http://www.genomicglossaries.com/
Life Science Dictionary from BioTech	http://biotech.icmb.utexas.edu/search/dict-search.html
Glossary from CancerPage.com	http://www.cancerpage.com/glossary/
Bioremediation Glossary from Natural and Accelerated Bioremediation Research (NABIR)	http://www.lbl.gov/NABIR/generalinfo/glossary.html
Multilingual Medical Terms	http://allserv.rug.ac.be/%7Ervdstich/eugloss/language.html
The Dictionary of Cell and Molecular Biology	http://www.mblab.gla.ac.uk/dictionary/

Solutions to Self-Test Quizzes

[2-1] b
[2-2] c
[2-3] a
[2-4] d
[2-5] a
[2-6] a
[2-7] c
[2-8] d
[2-9] c
[2-10] c

[3-1] asparagine N
glutamine Q
tryptophan W
tyrosine Y
phenylalanine F
[3-2] a
[3-3] d
[3-4] c
[3-5] d
[3-6] a
[3-7] c
[3-8] false
[3-9] c
[3-10] d

[4-1] d
[4-2] c
[4-3] a
[4-4] blastp d
blastn a
blastx c
tblastn b
tblastx e
[4-5] c
[4-6] a
[4-7] a
[4-8] b
[4-9] b
[4-10] c

[5-1] a
[5-2] b
[5-3] b
[5-4] c
[5-5] b
[5-6] a
[5-7] a
[5-8] a
[5-9] b
[5-10] d

[6-1] c
[6-2] c
[6-3] d
[6-4] a
[6-5] c
[6-6] c
[6-7] a
[6-8] b
[6-9] b
[6-10] c

[7-1] c
[7-2] c
[7-3] a
[7-4] b
[7-5] d
[7-6] d
[7-7] a
[7-8] d
[7-9] a

[8-1] a
[8-2] c
[8-3] b
[8-4] c
[8-5] c
[8-6] b
[8-7] d
[8-8] b
[8-9] a
[8-10] c

[9-1] a
[9-2] c
[9-3] d
[9-4] c
[9-5] d
[9-6] c
[9-7] a
[9-8] b
[9-9] d

[10-1] b
[10-2] d
[10-3] c
[10-4] b
[10-5] b
[10-6] b
[10-7] c
[10-8] a
[10-9] a
[11-1] d
[11-2] b
[11-3] c
[11-4] a
[11-5] b
[11-6] a
[11-7] b
[11-8] a
[11-9] a
[11-10] c

[12-1] c
[12-2] a
[12-3] d
[12-4] c
[12-5] a
[12-6] a
[12-7] d

[13-1] c
[13-2] d
[13-3] d
[13-4] a
[13-5] d
[13-6] b
[13-7] d

[14-1] b
[14-2] d
[14-3] a
[14-4] c
[14-5] c
[14-6] d
[14-7] a

[15-1] a
[15-2] c
[15-3] c
[15-4] c
[15-5] b
[15-6] d
[15-7] c
[15-8] b
[15-9] a
[15-10] a

[16-1] c
[16-2] c
[16-3] b
[16-4] d
[16-5] a
[16-6] a
[16-7] a
[16-8] a
[16-9] c

[17-1] c
[17-2] c
[17-3] a
[17-4] a
[17-5] a
[17-6] b
[17-7] b
[17-8] d

[18-1] a
[18-2] a
[18-3] b
[18-4] c
[18-5] a
[18-6] b
[18-7] c
[18-8] a

SUBJECT INDEX

Ab initio methods
  prokaryotic genome analysis, 481
  protein structure, 282, 309–310
ABC protein. *See* ATP-binding cassette.
ABCD1 gene, 351
Accepted point mutations. *See* PAM
Accession numbers, 26–27
Acquired immunodeficiency syndrome (AIDS). *See* Human immunodeficiency virus; Disease, human
Acrocentric chromosome. *See* chromosome
Actin, 164, 165, 192
β-Adrenergic receptor, 112–113, 233
Adrenocorticotropin, 40
Adrenoleukodystrophy, 351
Affinity chromatography, high-throughput protein analysis, 252–254
Affymetrix, 177, 179
Agglomerative hierarchical clustering, microarray data analysis, 204–210
Agriculture, genome sequencing applications, 419
Albumin, 222
Algorithms. *See* Basic Local Alignment Search Tool (BLAST); Heuristic search; Microarrays, data analysis; Multiple sequence alignment
  defined, 46
  dynamic programming, 63
  eukaryotic gene-finding algorithms, 428–429, 553–562
  FASTA search algorithm, 71–72, 134–135
    available in Protein Data Bank, 280, 289
  Feng-Doolittle progressive sequence alignment, 321–325
  pairwise sequence alignment, 46, 62–73
  unweighted pair-group method with arithmetic mean (UPGMA), 206, 323, 377, 379–382
Allelic variants
  complex genetic disorders, 669–673
  defined, 29, 660, 663
  locus-specific databases, 663–664
  Online Mendelian Inheritance in Man (OMIM) data, 660
  Single-nucleotide polymorphisms, 666–669
Allopolyploidy, chromosome organization, 565
α—helices, secondary protein structure, 275–281
Alternative splicing, RNA, 160
Alu repeat sequences
  eukaryotic genome sequencing, 547, 554
  Human Genome Project data, 608
Alveolates, eukaryotic genome sequencing, 570, 572–574
Amino acids
  genetic code, 54
  normalized frequencies, 53
  protein structure, primary-secondary transitions, 278–280
  relative mutabilities, 51, 53
  structures and one- and three-letter abbreviations, 45
  substitutions, molecular clock hypothesis, 360–363
  substitutions in human disease, 679, 683
Amyloid precursor protein, 222, 312
Amyotrophic lateral sclerosis (Lou Gehrig's disease). *See* Disease, human
Analysis of variance (ANOVA), microarray data, 200
Aneuploidies, chromosomal disease, 673
Angiosperms, phylogenetic analysis, 576–577
Animal genomes. *See* Genomes, eukaryotes.
Annotation techniques
  bacterial genomes, 428
  Drosophila genome sequencing, 586
  eukaryotic genomes, 428–429
    protein-coding genes, 553–562
  genomic DNA, 425–429
  human genome data, University of California Santa Cruz Bioinformatics database, 614–616
  prokaryotic genome analysis, 480–483, 486–488
  rice genome, *Saccharomyces cerevisiae,* 506–508
Anopheles gambiae (mosquito), 417, 566, 587
Aphrodisin, 7, 82
Apicomplexae, 572. *See* Genomes, eukaryotic, protozoan; *Plasmodium*
Apicoplast, malaria genetics and genomics, 573–574
Apolipoprotein D, properties, 7
Apoptosis
  protein function, 246
  viral genomes, 444
Arabidopsis thaliana
  C value paradox, 543
  genome sequencing, 415–416, 576–579
  human disease genes, 677
  mitochondrial genome, 407
  simple sequence repeats, 548
  transposable elements, 546
Archaea. *See* Genomes, archaea and bacteria
  defined, 399
Arthropods, genome sequencing, 585–587
Aspergillus fumigatus, genomic analysis, 528, 529
ATP-binding cassette (ABC) domain, 37–38, 287, 534
Autism, 649, 671
Automated database curation, databases, 341
Average-linkage clustering, microarray data analysis, 206
Avidin, 295

Bacteria. *See* Genomes, archaea and bacteria
Bacterial artificial chromosome (BAC)
  human genome sequencing, 620–621, 623
  identification using Grail software, 559
  libraries, shotgun sequencing, 421–422

Bacterial transposons, protein localization, 243
Bacteriophage ϕX174, sequencing history, 406–407
Bait proteins, high-throughput protein analysis
  affinity chromatography/mass spectrometry, 252–254
  yeast two-hybrid system, 254–256
BAliBASE algorithm, multiple sequence alignments, 347–350
Basic Local Alignment Search Tool (BLAST) 25, 86–154
  advanced searching
    alignment tools, genomic DNA searches, 135–137
    BLAST-like algorithms, 133–135
    BLAST-like tool (BLAT), 136–137, 152
    Ensembl BLAST, 130–131
    gene discovery, 147–150
    immunoglobulin BLAST (IgBLAST), 133
    Megablast algorithm, 136
    molecule-specific sites, 133, 169
    organism-specific sites, 128–133
    pattern-hit initiated BLAST (PHI-BLAST), 145–147
    position-specific iterated BLAST (PSI-BLAST), 137–144
      corruption errors, 144
      performance assessment, 143–144
    position-specific scoring matrix (PSSM), 138–145
    specialized servers and algorithms, 127–135
    TIGR BLAST, 130–133
  algorithm, 100–108
    E and *P* value relations, 106–107
    Karlin-Altschul statistics, 105–106
    list, *scan*, *extend* processes, 101–103
    masking repetitive and low-complexity sequence, 94–96, 551
    raw score/bit score data, 106
  statistics and *E* value, 103–106
  basic properties, 87–89
  BLAST 2 Sequences, 72–74
  Composition-based statistics, 94, 139
  database selection, 92, 119
  expect (E) value, 77
    BLAST output, 98
    defined, 96–97
    equation, 105–107
    interpretation of BLAST results, 109–113, 117–119, 122–123
    PSI-BLAST search, 138–142, 144
    relation to probability value *P*, 107
  filtering sequences, 94–96
  formatting options, 97–100
  gapped alignments, 107
  high-scoring segment pairs (HSPs), defined, 77, 88
  high-scoring segment pairs (HSPs), BLAST algorithm, 102–108
  human genome data, Ensembl human genome browser, 130–131, 612
  program selection (blastp, blastn, blastx, tblastn, tblastx), 90–92
    gene discovery, 147–150
  protein domain/motifs, 227–228
  protein structure, 287, 289
  input (query), 89–90
  Saccharomyces cerevisiae (budding yeast), genome duplication, 515
  search parameter selection, 92–100
  search steps, 89–100
  search strategies
    general concepts, 108
    multidomain protein query, 114–117
    results evaluation, 108–113
    scoring matrices, 117–122
  uses, 87–88
  viral genomes, 449, 452, 453, 458
  word size, 97
  WU BLAST, specialized algorithms, 134–135
Basidiomycetes, genomic analysis, 505, 530
Beetles, 538
Berkeley Drosophila Genome Project (BDGP)
  Drosophila genome sequencing, 586
  fly chromosome sequencing history, 415
BestFit program, pairwise sequence alignment, 71, 701–706
β sheets, secondary protein structure, 275–281
Between-subject design, microarray data analysis, inferential statistics, 200–201
Bilateria, genome sequencing, 582–583
Biochemical pathways, protein function, 245–246, 258–264
Bioinformatics
  defined, 3–4
  key websites, 10–11
  overview, 3–7
  research perspectives, 4–7
Biological databases. *See* databases
Bit scores, Basic Local Alignment Search Tool (BLAST), 106
BLAST. *See* Basic Local Alignment Search Tool (BLAST)
BLAST 2 Sequences, 72–74
BLAST-like alignment tool (BLAT)
  advanced searching, 134–137
  compared to BLAST, 152
  mouse genome analysis, 589
BLOCKS database, multiple sequence alignment, 60, 332, 334
Blocks substitution matrix (BLOSUM)
  human disease models, 679, 683
  matrices in Basic Local Alignment Search Tool (BLAST), 97
  pairwise sequence alignment, 60–61
  position-specific iterated BLAST (PSI-BLAST), 138–139
BLOSUM. *See* Blocks substitution matrix
Bonferroni correction, 76, 165, 199
Bootstrapping, tree evaluation, 386–389
Brain, 157, 166, 180
Budding yeast. *See Saccharomyces cerevisiae* (budding yeast)

Caenorhabditis elegans. *See* Genomes, eukaryotic, metazoa
  genome sequencing, 414, 583–585
  genome size, 544
  human disease genes, 677
  simple sequence repeats, 547
  transposable elements, 546
Calycin superfamily, 295
Cambrian explosion, genomic analysis, 401, 402, 583
Cancer. *See* Disease, human
Candida albicans, genomic analysis, 529
Capillary electrophoresis sequencing, human genome sequencing, 621
Case Western Reserve University (CWRU) Duplication Browser, eukaryotic genome sequencing, 547
CATH database, protein structure, 297–301
cDNA (complementary DNA) library, 162–168
  biases and normalization, 167–168
  eukaryotic genomes, 556–557
  expressed sequence tags (ESTs), 20–22, 163
  human genome genes, 635
  mouse orthologs of human disease genes, 684
  yeast two-hybrid system, 254–256
Celera Genomics
  database characteristics, 16
  human genome data, 608, 609
    whole-genome shotgun (WGS) sequencing, 620
Cell, bioinformatics and role of, 4–6
Cellular pathways, 258–262
Cellular processes, protein function, 243–247
Centromere
  eukaryotic genomes
    chromosome organization, 543, 549–550, 564–566, 597
    in yeast, 509, 516, 517
    tandem repeats, 549–550
  human genome, 633
Character-based methods, tree-building. *See* Molecular phylogeny
Characteristic value u. *See* Extreme value distribution
Chime visualization program, protein structure, 292, 295
Chitin, 504

Chloroplast
genome sequencing history, 408, 410
plant genome sequencing, 576
Chou-Fasman algorithm, protein structure, 279–280
Chromatin diminution, eukaryotic chromosomes, 566
Chromosomal disorders, 673–674
Chromosome. *See also* Centromere; Chromatin diminution
acrocentric chromosome, defined, 569
allopolyploidy, chromosome organization, 565
conserved synteny map, 621
databases, 565
fragile sites, eukaryotic chromosomes, 566
fused, 566
human chromosome 2, fused chromosome organization, 566
human chromosome 19, 636
human chromosome 21, genomic analysis, 517, 566
human, size classification, 626–627
inverted chromosome regions, basic properties, 566
metacentric chromosome, defined, 565
metaphase, chromosome analysis, 564
neocentromere, 550
Saccharomyces cerevisiae, typical chromosome, 508–511
sequencing history, first eukaryotic chromosome, 408–409
web resources, 565
Ciona intestinalis (sea squirt), genome sequencing, 587–588
Clade, molecular phylogeny, 367
ClustalW, multiple sequence alignment
description, 343–344
example, 321–326
use in phylogenetic analysis, 376
Cluster analysis. *See* Microarrays, data analysis
Clusters of Orthologous Groups (COG) database
bacteria/archaea genome classification, 486–489
used in Entrez descriptions of genomes, 410–411
protein function, 246–247
tree-building, 377
Cn3D software, protein structure, 291
Cnidaria, 583
Codons
database resources, 563
gene annotation in *Saccharomyces cerevisiae*, 507
Coelacanth, 82
Coelomata (bilaterian animals), genome sequencing, 583
COG. *See* Clusters of Orthologous Groups
Coiled-coil domains, 236–238
Comparative modeling. *See* Protein structure
Complementary DNA library. *See* cDNA library
Complex genetic disorders, 669–673
Composition-based statistics. *See* Basic Local Alignment Search Tool (BLAST)
Conservative substitutions, pairwise sequence alignment, 46–47
Conserved Domain Database (CDD)
as option in Basic Local Alignment Search Tool (BLAST), 93
multiple sequence alignment, 330, 339–340
Conserved synteny, 567
Contig map
Human genome data
Ensembl project, 612–613
National Center for Biotechnology Information (NCBI), 623, 625–627
Convention on Biological Diversity, taxonomy and bioinformatics, 404
Corruption errors, position-specific iterated BLAST (PSI-BLAST), 144
Coypu, 359, 360
CpG islands
human genome sequencing, 627–629
identification, 559
Critical Assessment of Techniques for Protein Structure Prediction (CASP), 311
Critica program, prokaryotic genome annotation, 480–481
Crustacyanin, 268
Cryptomonads, genomic sequencing, 416–417, 576
Curse of dimensionality, microarray data analysis, 203, 211
C value paradox, eukaryotic genomes, 543–544
Cyclic adenosine monophosphate (cAMP), 112, 582
Cystic fibrosis transmembrane regulator (CFTR), 312, 314, 677
Cytochrome *c*, 285, 318, 360, 362

Dali Domain Dictionary, protein structure, 301–304
Danio rerio (zebrafish), genome sequencing, 589
Databases and data repositories. *See also* specific databases; European Bioinformatics Institute; GenBank; National Center for Biotechnology Information; Websites
disease, 659–666
DNA compared to gene expression, 160
genome sequencing data, 425–427
microarray data, 182, 184
Dayhoff model. *See* Pairwise sequence alignment; PAM
dbSNP database, single-nucleotide polymorphisms (SNPs), 668–670
Deletions. *See* Chromosome
genomic deletions, yeast strains, 522–528
pairwise sequence alignment, 47
Descriptive statistics, microarray data analysis. *See* Microarrays, data analysis
Dialign program, multiple sequence alignments, 345, 348
Dictyostelium discoideum (slime mold; amoeba)
genome sequencing, 582
genome size of amoebae, 420
novel lipocalin gene, 149–150
Dideoxynucleotide sequencing
human genome data, 621
defined, 406–407
Diploid organisms
eukaryotic chromosome organization, 541, 543, 564
genome size and, 467
yeast strains, 526
Diplomonads (diplomonadida), eukaryotic genome analysis, 570–571
Disability-adjusted life years (DALYs), global burden of disease, 654
Disease
amino acid substitutions, 679, 683
animal diseases
bovine spongiform encephalophathy ("mad cow" disease), 312, 438
murine acquired immunodeficiency syndrome (AIDS), 684
Online Mendelian Inheritance in Animals (OMIA), 677
rinderpest, 453
Simian Immunodeficiency Virus (SIV), 114, 447–452
bacterial genome classification, 473–474
bioinformatics perspective, 649–650
categories, 652–653
chromosomal abnormalities, 565–566, 673–674
classification, 653–655
complex disorders, 669–673
databases
dbSNP at NCBI, 669
GeneCards, 661, 665
Human Gene Mutation Database, 661
mutation, 661–666
Online Mendelian Inheritance in Man (OMIM), 659–664
organellar and pathway, 675–676
defined, 648

Disease (*cont.*)
DNA variation, 647–649
environmental, 652–653
functional disease gene classification, 684–687
Garrod's theses, 650–651
genomic sequencing criterion, 419
global burden, 654
human disease organizations, 684
human diseases
acquired immunodeficiency syndrome (AIDS), 7, 446–453. *See* Human immunodeficiency virus
adrenoleukodystrophy, 351, 570, 656
albinism, 651
alkaptonuria, 650–651
Alzheimer disease, 312
amyotrophic lateral sclerosis (Lou Gehrig's disease), 260, 264
anthrax, 474
autism, 649, 671
cancer, 37, 681, 689
Chagas' disease, 571
Cruetzfeld-Jakob disease, 438
cystic fibrosis, 311–312, 314, 656
cystinuria, 651
deletion 11q syndrome, 566
diabetes, 669, 685
Down syndrome, 517, 566
fungal disease, 502, 529
hemophilia, 566, 656
Huntington's disease, 153, 547
infectious disease, 441, 473–474, 651–656, 684
influenza, 439, 440
Kaposi's sarcoma, 443–446
lead poisoning, 207–210, 653
leishmaniasis, 572
leprosy, 469, 474
Lyme disease, 466, 474
malaria, 572–574, 684
measles, 453–455
mosquito-borne, 587
phenylketonuria, 652
pneumonia, 473, 489, 684
Rett syndrome, 51, 180, 227, 656–659
ringworm (tinea), 503
salmonellosis, 474
schizophrenia, 116, 649
sickle cell anemia, 312, 314, 652, 656–657
sleeping sickness, 571
smallpox, 438
thrush (oral candidiasis), 503
trachoma, 489
trisomy 13, 673
trisomy 18, 673
trisomy from neocentromeres, 550
tuberculosis, 474, 474
uniparental disomy, 566
viral diseases, vaccine-preventable, 441
viral hepatitis, 684
Wilms tumor, 689
incidence, 447
model organisms, 676–684
monogenic disorders, 655–669
mutation databases, 659–667
positionally cloned genes, 686
prevalence, 447
protein structure, 311–312, 314
single nucleotide polymorphisms, 666–669
vaccine-preventable, bacterial, 474
vaccine-preventable, viral, 441
websites on human disease, 675, 686
Distance-based methods, tree-building. *See* Molecular phylogeny
Distance measures *vs.* similarity, multiple sequence alignment, 326
Divisive hierarchical clustering. *See* microarray data analysis
DNA versus protein, pairwise sequence alignment, 41–42
DNA (deoxyribonucleic acid)
accession numbers, 26–27
bioinformatics research and, 3–4
biological databases on, 15–16
accession protocols, 27–33
GenBank database entries for, 17–24
genome annotation, 425–429
genome complexity, reassociation rates, 544–545
genome sequencing, 421–425
phylogenetic analysis, 371–374
sequencing. *See* Dideoxynucleotide sequencing
DNA microarrays. *See* Microarrays
Domain Architecture Retrieval Tool (DART), multiple sequence alignment, 333, 340
Domains. *See also* Protein, domains and motifs
defined, 225–226
Dot plots, pairwise sequence alignment, 73, 75
Down syndrome cell adhesion molecule (DSCAM)
alternative splicing, 160
Anopheles gambiae ortholog, 587
Drosophila melanogaster (fruitfly)
C value, 543–544
genome sequencing, 585–587
human disease models, 677–680
sequencing history, 415, 416, 422
Drosophila virilis, microsatellites, 547
DSCAM. *See* Down syndrome cell adhesion molecule
DUST program, repetitive repetitive DNA analysis, 551
Dye ratios, microarray data analysis, 190–192
Dysmorphic Human-Mouse Homology Database, mouse disease models, 683
EBI. *See* European Bioinformatics Institute
EcoCyc/MetaCyc database, cellular pathways, 259–262
Edman degradation, protein sequencing, 249–250
Eigenvalues, principal components analysis, 212
Eigenvectors, principal components analysis, 212
Electronic Northern blot, cDNA libraries, 164, 167
Electronic polymerase chain reaction (E-PCR) sequence tag sites, 560
Elongation factors, phylogenetic analysis, 478, 504, 540
EMBL/EBI database, 10, 31
proteome tools, 51
specialized database search algorithms, 134–135
viral genome resources, 455
Encephalitozoon cuniculi, 529
C value, 543
eukaryotic genome sequencing, 570
endocytosis, 541
Ensembl project, 31–32, 37
annotation pipeline, 425, 427
advanced database search techniques, 130
distributed annotation system, 581
genome browser for *Fugu*, 588
genome browser for mosquito, 587
genome browser for mouse, 589
Human Genome Central, 608
human genome data, 609
access points, 610, 612–621
browser front page, 614
Entrez database, 24, 31
used to limit Basic Local Alignment Search Tool (BLAST) search, 93–94, 113
genome site, 405–409
pol (polymerase) gene example, 33–34
prokaryotic genome analysis, TaxPlot, 493
Protein Data Bank (PDB) entries, 289–292
Saccharomyces cerevisiae chromosome analysis, 508, 510–511
viral genomes, 452, 457–458
Environmental disease. *See* Disease
Enzyme Commission (EC) system, protein function, 246
Epicellular bacteria, genomic classification, 471–473
Epidermal growth factor receptor, 113
Epifagus virginiana (beech drop), 575
Epitope tags
exogenous transposon harnessing, 523–526
protein localization, 243
ESTs. *See* Expressed sequence tags

Ethical, Legal and Social Issues (ELSI) initiative, Human Genome Project, 619
Euchromatin, defined, 545, 617
Euclidean distance, microarray data analysis, 203
Eukaryotic chromosome. *See* chromosome
Eukaryotic Gene Orthologs (EGO) (TIGR), gene expression bioinformatics, 169
Eukaryotic genomes. *See* Genomes, eukaryotic
European Bioinformatics Institute, 10
  accession protocols, 31–32
  ARTEMIS, 493
  prokaryotic genome analysis, 493
  proteome tools for human, 636, 638–640
  proteome tools for *Saccharomyces cerevisiae*, 506, 508–509
European Small Subunit Ribosomal RNA Database, molecular phylogeny, 401
Evidence codes, Gene Ontology (GO) Consortium, 240
Evolution. *See* molecular phylogeny.
  defined, 357–358
  disease, 647
  pairwise sequence alignment, homology and, 47–48
Exons
  eukaryotic genomes, 135, 553–563
  size distribution, species variation, 636–637
Expect (E) value. *See* Basic Local Alignment Search Tool (BLAST)
Expert Protein Analysis System (ExPASy), 33, 230
  protein analysis website, 230–231
    pathway maps, 257, 259
    physical properties, 235
    two-dimensional gel electrophoresis, 251–254
  protein structure, 292, 294
Expressed sequence tags (ESTs)
  Basic Local Alignment Search Tool (BLAST), 88
    gene discoveries, 147–150
    pol protein searches, 116–117
    TIGR BLAST, 130
  eukaryotic genomes, 555–557
  GenBank division, 20–21, 26
  gene expression
    cDNA libraries, 162–168
    TIGR indices, 169
  genome sequencing, 429–430
  single-nucleotide polymorphisms (SNPs), 666, 668–669
Extreme value distribution (EVD). *See* Basic Local Alignment Search Tool.
  characteristic value u, 104–106
  defined, 103–105
  Hidden Markov Models and HMMER, 328
  historical description, 86
Extremophiles, bacteria/archaea genomes, 472–473
Extrinsic algorithms, eukaryotic genomes, protein-coding genes, 429–430, 555–560

False discovery rate, Significance Analysis of Microarrays (SAM), 201–202
Family, defined, 225–229. *See* Protein
FANTOM database, mouse disease models, 681, 684
FASTA (FAST-All) search algorithm, 71–72, 135
  available in Protein Data Bank, 280, 289
FASTA sequence format, 31–32
  used as input for BLAST searches, 89, 130, 132
  used as input for multiple sequence alignments, 321
  used as input for genomic DNA analysis algorithms, 550–551, 557
Fatty acid-binding proteins (FABPs), SCOP database, 295
Feng-Doolittle progressive sequence alignment, 321–325, 365
Fertility, chromosomal abnormalities and, 673–674
Fibrinopeptides, 360, 362
Fibronectin repeat, 227–229
Filtering. *See* Basic Local Alignment Search Tool (BLAST)
Find-a-gene exercise, 9–10
Fingerprints. *See* Motifs
Fisher's exact test, gene expression bioinformatics, cDNA libraries, 165–167
Fish genomes, sequencing, 588–589
Fission yeast. *See Schizosaccharomyces pombe* (fission yeast)
FlyBase, *Drosophila* genome sequencing, 586
Folds, defined, 227
Footprinting, transposon techniques, 522–523
Fragile sites, eukaryotic chromosomes, 566
Free journal access, debate, 35–36
FSSP database, protein structure, 302, 305–307
Fugu rubripes (pufferfish)
  C value, 543
  Fugu genome browser, 588–589
  genome sequencing, 588–589
Functional genomics. *See* Microarrays
  defined, 3–4, 174, 246
  fungi, 520–528. *See* Fungi, functional genomics
  nematode, 584–585
Fungi. *See* Genomes, eukaryotic; *Saccharomyces cerevisiae* (budding yeast)
  description and classification, 503–504
  functional genomics, 520–528
    exogenous transposons, 523–526
    genetic footprinting with transposons, 522–523
    molecular barcoding, genomic deletions, 526–528
    sporulation studied with microarrays, 193, 194
  human disease models, 677–679, 681–682

GAL4 domains, yeast two-hybrid system, 254–256
Gaps in sequence alignment, 47
  Basic Local Alignment Search Tool (BLAST), 107
  phylogenetic analysis, 375–377
GAP program
  pairwise sequence alignment, 44, 46, 67–69, 100, 701, 703–706
Garrod's hypothesis, genetic disease, 650–651
GC content (guanosine plus cytosine)
  human genome
    distribution analysis, 638
    long-range variations, 627–628
    protein-coding genes, 636
  prokaryotic, 478, 480
  eukaryotic, 509, 543, 563, 580–581
GCG. *See* Genetics Computer Group
GenBank database, 16–24
  data capacity, 16–17
  data categories and classification, 19–20
  expressed sequence tags (ESTs), 20–21
  genome survey sequences (GSSs), 22–24
  high-throughput genomic sequence (HTGS), 23
  HTGS division, finished/unfinished data, 422–423
  organisms and species represented, 17–19, 22
  sequence-tagged sites (STSs), 22
  UniGene project, 21–22, 31
  viral genomes, human immunodeficiency virus, 451–452
Gene
  Basic Local Alignment Search Tool (BLAST) gene discovery, 147–150
  bioinformatics research and, 3–7
  biological databases on, 10, *15–16*
  definition, 551–553
  disease genes, functional classification, 684–688
  human genome, 633–638
  open reading frame (ORF), defined, 506

Gene (*cont.*)
prediction in eukaryotic genomes, 428–430, 553–563
prediction in prokaryotic genomes, 428, 480–483
GeneCards, monogenic disease data, 661, 665
Gene conversion, 511
Gene duplication, *Saccharomyces cerevisiae* duplication and gene acquisition, 511–519
Gene expression. *See* microarrays
bacterial, 156
basic principles, 157–162
cDNA libraries, 162–169
limits and pitfalls, 166–168
correlation with protein expression, 174–176
microarrays, genomewide measurement, 172–183
biological confirmation, 182
data analysis, 180–182
databases, 182
defined, 172–176
experimental design, 177–178
hybridization, 178–179
image analysis, 179–180
RNA/probe preparation, 178
websites, microarray data repositories, 184
mRNA, 160–162
serial analysis of gene expression (SAGE), 169–172
TIGR gene indices, 169
Gene Ontology (GO) Consortium
evidence codes, 240
protein analysis
biological process, defined, 246
conceptual framework, 224
databases and resources, 237–241
proteome analysis, human genome, 636, 640–641
Gene/protein trees, structure and properties. *See* Molecular phylogeny
Gene silencing, human genome, 628
Genetic code, defined, 54
Genetic mapping, human genome sequencing, *vs.* physical mapping, 629–630
Genetics Computer Group (GCG)
description, 68
general commands, 697–699
multiple sequence alignment, PileUp, 345–347, 709–712
pairwise alignment, 44, 68, 701–709
phylogenetic analysis, 712–713
sequence analysis, 713–716
sequence entering and editing, 699–701
Génolevures project, fungi, 516–517
Genome Analysis Pipeline, gene-finding algorithms, 558–562
Genome duplication, 511–519, 578–579, 642
Genome sequencing
annotation, 425–429
annotation pipeline, 425, 427
cost, 419
coverage, 424–425
finished/unfinished project status, 422–424
high throughput genomic sequences, 423
historical overview, 404
human genome features, 623, 625–626
sequencing centers, 421
size, 420–421
strategies, 421–422
trace archives, 425–426
Genome survey sequences (GSSs), GenBank division, 22–24
Genomes. *See* Tree of life
bioinformatics and, 3–7
chronology, 404–417
completed projects, websites for, 405
complexity, reassociation rates, 544–545
defined, 397–398
genomes, archaea and bacteria
classification, genome size and geometry, 467–469
classification, human disease relevance, 473–474
classification, lifestyle, 470–473
classification, molecular sequences, 475–478
classification, morphology, 466–467
classification, ribosomal RNA sequences, 474–475
comparison to eukaryotes, 541–543
gene expression, 156
genomic analysis, 478–495
species, archaea
annotation and comparison, 486–495
Archaeoglobus fulgidus, 414
COG comparisons, 486–489
gene-finding, 480–483
lateral gene transfer, 483–486
Methanococcus jannaschii, 413, 473
MUMmer, 490–495
noncoding and repetitive DNA sequences, 543
nucleotide composition, 478–480
Sulfolobus solfataricus, 415
TaxPlot, 489–490, 493
Thermoplasma acidophilum, 415
species, bacteria
Bacillus anthracis, 490
Borrelia burgdorferi, 466
Buchnera aphidicola, 473
Caulobacter crescentus, 251
Chlamydia, 414, 415, 489–490
Escherichia coli, 413, 471, 472, 493, 496, 543
Haemophilus influenzae, 15, 409–412, 421
Helicobacter pylori, 256, 260, 474
Mycobacterium leprae, 469
Mycobacterium tuberculosis, 472, 480
Mycoplasma genitalium, 413, 421, 437, 469, 471
Mycoplasma pneumoniae, 473
Nanoarchaeum equitans, 437, 469
Rickettsia prowazekii, 414, 473, 496
Selenomonas sputigena, 465
Streptomyces coelicolor, 421, 467
websites, 496
genomes, eukaryotic
chromosomes, 564–566
compared to prokaryotes, 541–543
comparison of genomic DNA, 566–567
C value paradox, 543
finding genes, 553–562
noncoding and repetitive DNA, 543–553
overview, 539–541
paradox of protein-coding genes, 562–563
transcription factor databases, 563–564
genomes, eukaryotic, protozoan and various other
Dictyostelium discoideum (amoeba), 420, 582
Giardia lamblia, 540, 549, 570–571
Guillardia theta, 576
Leishmania major, 572
Paramecium, 549, 574
Plasmodium falciparum, 417, 424, 549, 572–574
Plasmodium yoelii yoelii, 417, 549, 573–574
Tetrahymena, 549, 574
Trichomonas, 570, 571
Trypanosoma, 571–572, 597
genomes, eukaryotic, fungi, 503–537. *See* Fungi
Aspergillus fumigatus, 528, 529
basidiomycetes, 530
Candida albicans, 502, 528, 529
Kluyveromyces lactis, 517
microsporidia, 529, 570
Neurospora crassa, 528, 529–530
Saccharomyces bayanus, 517
Saccharomyces cerevisiae (budding yeast), 408–409, 413, 505–519, 677
Schizosaccharomyces pombe (fission yeast), 417, 528, 530–531, 573, 677
genomes, eukaryotic, metazoa (animals)
Anopheles gambiae (mosquito), 417, 566, 587
Ascaris suum, 549, 566
Caenorhabditis elegans (nematode), 414, 583–585, 677

Ciona intestinalis (sea squirt), 587–588
Danio rerio (zebrafish), 589
Drosophila melanogaster (fruitfly), 415, 416, 422, 566, 585–587, 677
Fugu rubripes (pufferfish), 556, 588–589
Homo sapiens (human), 414, 549. *See also* Human genome
Mus musculus (mouse), 589–591, 679–685
primates, 419, 591–593, 595
genomes, eukaryotic, plants. *See* Plant genomes
Arabidopsis thaliana (thale cress), 407, 415, 416, 576–579, 677
Epifagus virginiana (beech drop), 575
Oryza sativa (rice), 560–562, 579–581
Zea mays (maize), 546
genomes, organelles
chloroplast, 408
mitochondrion, 407–408
genomes, prokaryotes 466–500
annotation and comparison, 486–489
COG comparisons, 486–489
gene identification, 480–483
GLIMMER, 481, 482
lateral gene transfer, 483–486
minimum size, 471
MUMmer, 490–495
nucleotide composition and GC content, 478, 480
Tax Plot, 489–490, 493
genomes, virus
diversity and evolution, 441–442
Los Alamos National Laboratory (LANL) database, 453–456
National Center for Biotechnology Information (NCBI) resources, 452
sequencing history, 406–407
species
bacteriophage ϕX174, 406–407
herpesvirus, 442–446
human immunodeficiency virus (HIV), 439, 442, 446–453, 461
lentivirus, 446, 451
measles virus, 453–455, 457–458
mimivirus, 438
poxviruses, 459
simian immunodeficiency virus (SIV), 447, 449, 451, 452
virus classification, 438–439
web resources, 460
GENSCAN program, eukaryotic gene predictions, 559, 561
Geological history. *See* History of life on earth
Giardia lamblia
eukaryotic genome sequencing, 570–571
GIRI Censor Server, repetitive DNA, 550, 552
GLIMMER program, 428–429, 481–483, 557
Global sequence alignment, 62–68
Feng-Doolittle progressive sequence alignment, 321–325
pairwise sequence alignment
basic principles, 62–63
Genetics Computer Group (GCG) GAP program, 68, 701–704
Needleman and Wunsch algorithm, 63–68
statistical significance tests, 76–77
web-based algorithms, 79
Global normalization, microarray data analysis, 192–194
local normalization *vs.*, 194–197
Globin genes and proteins
disease, 312, 314, 647, 657
evolution, 43, 50, 80, 360, 362
gene expression, 165
structure, 222, 272, 275, 288, 296
Glutamate, 257
Glutamate dehydrogenase, 50, 362
Glyceraldehyde 3-phosphate dehydrogenase (GAPDH)
gene assigned to different chromosomes, 547
multiple sequence alignment, 48, 53, 320
used to normalize gene expression values, 192
used to make tree of life, 431
Glycodelin, 38
Glycosylation sites, websites, 265
G proteins, viral genomes, 444, 449
GrailEXP program, gene-finding predictions, 557–563
GRAIL program. annotation, genomic DNA, 425, 429
Guanosine triphosphate (GTP)-binding proteins, domains and motifs, 226–227
G value paradox, eukaryotic protein-coding genes, 562–563
Gyrase, 256

Haemophilus influenzae. *See* Genomes, archaea and bacteria
Hamming distance, tree-building, 378–379
Haploid chromosomes
genome size and, 467
eukaryotes, 541
prokaryotes, 543
yeast, 526
Haplotype
defined, 587
mosquito genome, 587
single nucleotide polymorphisms (SNPs), 668
HapMap project, 668
HapMap project, single-nucleotide polymorphisms (SNPs), 668
Hemagglutinin (HA). *See* Epitope tag
Hemoglobin. *See* Globin
Herpesvirus, herpesviruses
microarray analyses, 445–446, 450
phylogeny and gene expression, 442–446
Heterochromatin, eukaryotic chromosomes, 545, 564, 566, 586
Heterogeneity, complex genetic disorders, 671
Heterotrophs, 504
Heuristic search
defined, 46
pairwise sequence alignment, FASTA and BLAST algorithms, 71–72
phylogenetic trees, 370, 384
Hidden Markov models (HMMs)
multiple sequence alignment, 325–329
Pfam database, 331–332
prokaryotic genome analysis, 481
Hierarchical cluster analysis. *See* microarray data
Hierarchical shotgun sequencing, 421–423
human genome sequencing, 620–623
High-scoring segment pairs (HSPs). *See* Basic Local Alignment Search Tool (BLAST)
High-throughput genomic sequence (HTGS). *See* GenBank
Hippopotamus, 82
Histone, 50, 362
History of life on earth, 47–48, 399–403
HIV. *See* Human immunodeficiency virus.
HMMER program, multiple sequence alignment, hidden Markov models, 328–331
Homology
definition, 41–44
phylogenetic analysis, multiple sequence alignment, 375–377
protein function, 243–245
protein structure, 274–276
comparative modeling, 282, 303–309
within *Saccharomyces cerevisiae* (budding yeast) genome, 511–519
Homo sapiens. *See* Human genome
Housekeeping genes. *See* Glyceraldehyde 3-phosphate dehydrogenase
Human Gene Expression (HuGE) Index database, 192
human genome sequencing, CpG islands, 628–629
microarray data analysis, global normalization, 192
Human Gene Expression (HuGE) Index database, microarray data analysis, global normalization, 192
Human Gene Mutation Database (HGMD), monogenic disease data, 661

Human genome
assembly, 623–627
clone/sequence coverage, 623–626
comparative proteome analysis, 636
CpG islands, 628–629
data resources, 611–619
Ensembl website, 610–621
NCBI website, 609–610
UCSC Human Genome Browser, 614–616, 622–623
draft sequences, 414–415
GC content, long-range variation, 627–628
gene content, 633–640
genetic/physical distance comparisons, 629
hierarchical shotgun sequencing, 620–623
human-mouse homology map, 590, 683–684
noncoding genes, 634, 635
primate sequencing data, 591–593
protein-coding genes, 636
proteome complexity, 637–640
repeat content, 629–633
segmental duplications, 632–633
simple sequence repeats, 632
transposon-derived repeats, 630–632
sequence features, 623–626
sequencing centers, 625
sequencing strategies, 620–623
websites, 608, 609, 642
Human Genome Organisation (HUGO), monogenic disease data, 661, 666
Human Genome Project (HGP)
background, 618–619
Ethical, legal and social issues (ELSI), 619
goals summary, 624
major findings, 608–609
Human Genome Variation Society (HGVS) database, mutation data, 664, 667
Human immunodeficiency virus (HIV). *See Pol* (polymerase)
Bioinformatics approaches, 451–456
diversity and evolution, 446–451
genome sequencing, 420
HIV Drug Resistance Database, viral genome resources, 453
HIV protease database, Frederick Cancer Research and Development Center, 453
Human vacuolar protein-sorting protein 45, position-specific iterated BLAST (PSI-BLAST), 142
Hydroxyapatite chromatography, genomic DNA analysis, 545

ICTV database, viral classification, 438–440
Identity, aligned sequences
pairwise sequence alignment
definitions, 42–47
percent identity, 73–75
Identity matrix, 64
Ideogram
autism etiology, 672
human genome data, 610, 611, 614, 617, 620, 621
karyotypes, 564
IMAGE Consortium, gene expression, 169
Image analysis, microarrays, 179–180, 191–192
Immunoglobulin BLAST (IgBLAST), molecule-specific sites, 133
Immunoglobulin domains, 229
Inborn factors, genetic disease, 651
Incidence, defined, 447
Infectious disease
morbidity/mortality rates, 653–654
mouse susceptibility models, 684
Inferential statistics, microarray data analysis, 198–203
Inferred trees, defined, 365
Insect genomes
Anopheles gambiae, 587
Drosophila melanogaster, 585–587
Insertions, pairwise sequence alignment, 47
Insulin, 318, 358–360
Integrated Gene Index (IGI), human genome analysis, 636
Integrated Molecular Analysis of Genomes and Their Expression (IMAGE) Consortium, gene expression, 169
Interleukin 8 receptor, 444–445
International Human Genome Sequencing Consortium (IHGSC)
human genome data, 609
Human Genome Project report, 618–619
International Rice Genome Sequencing Project (IRGSP), rice sequencing data, 580
International Statistical Classification of Diseases and Related Health Problems (ICD), disease classification system, 654–655
InterPro database
multiple sequence alignment, 336, 339–340
protein family definitions, 226
proteome analysis, human genome, 636–640
website, 225
Interspersed repeats. *See* Repetitive DNA
Intracellular bacteria, genomic classification, 472–473
Intrinsic algorithms, eukaryotic genomes, protein-coding genes, 429–430, 555–560
Introns
databases, 563
defined, 160
eukaryotic genomes, 135, 563, 582, 588
fungal, 506, 509
Saccharomyces cerevisiae chromosome analysis, 509
size distribution, species variation, 636, 637
iProClass database, multiple sequence alignment, 340
Isochores, human genome data, 628
Isoprenes, malaria genetics and genomics, 573–574
ISREC server, Coils program, 238
Iterative algorithms, multiple sequence alignments, 345, 348

Jackson Laboratory mouse website, mouse disease models, 683
mouse disease models, 683
mouse genome, 591–593
JalView, 332, 336–337, 365
Joint Genome Institute (Dept. of Energy), *Ciona intestinalis* sequencing, 587–588
Jukes-Cantor one-parameter model, tree-building, 379
Junk DNA, 545. *See* Repetitive DNA sequences

Kaposi's sarcoma-associated herpesvirus (HHV8), 443
Karlin-Altschul statistics, Basic Local Alignment Search Tool (BLAST), 105–106
Karyotype. *See* Ideogram
defined, 564
human karyotypes, 565
Kimura two-parameter model, tree-building, 379
Kinetochore, 550
Kinetoplast DNA, 572
k-means clustering, microarray data, 210
k-mers, human genome simple sequence repeats, 632
Krebs cycle, 245–246, 260
Kyoto Encyclopedia of Genes and Genomes (KEGG), cellular pathways, 260, 263–264

Labeling efficiency, microarray data analysis, 191–192
β-Lactoglobulin, 38. *See* Lipocalins
pairwise sequence alignment, 42–44, 46–47
structure, 293, 313, 320
Lateral gene transfer
description, 401–402, 483–486
Human Genome Project data, 608
pitfall in phylogeny, 570
LCN1 gene, eukaryotic genome sequencing, 547–549
Lead poisoning. *See* Disease, human

Leishmania species, eukaryotic genome sequencing, 572
Lipocalins. *See* also β-lactoglobulin, odorant-binding protein, retinol-binding protein
  eukaryotic genome sequencing, 535, 548, 549
  gene discovery, 147–150
  GXW motifs
    defined, 231–232
    multiple sequence alignment, 324, 335, 336
    pattern-hit initiated BLAST (PHI-BLAST), 145–147
    protein structure, homology and functional genomics, 274, 295
  input for Basic Local Alignment Search Tool (BLAST), 117–122
  multiple sequence alignment, 49
  multiple sequence alignment, hidden Markov models, 328–331
  pairwise sequence alignment, 42–47
    homology, 49–50
  pattern-hit initiated BLAST (PHI-BLAST), 145–147
  Protein Data Bank (PDB) queries, 288–289
  protein domains and motifs, 240, 486
  protein function, 243–246
  structural genomics, 275, 276, 313
Literature databases. See MEDLINE; PubMed
Local alignment algorithms
  GCG Bestfit and GAP programs, 701–706
  pairwise sequence alignment
    basic principles, 62–63
    Smith and Waterman algorithm, 69–71
    statistical significance tests, 77
Local normalization. *See* Microarrays, data analysis
LocusLink, 27–29, 31
  Human Map Viewer, 610, 642
  monogenic disease data, 660
  retinol-binding protein 4, 28–30, 161
Locus-specific databases, genetic disease, 661–667
Loess function, microarray data analysis, local normalization, 196–197
Log-odds score matrix, pairwise sequence alignment, 57–59
Log ratios, microarray data analysis, 192, 194
Long-branch attraction, maximum parsimony, 384, 386–387, 570
Long interspersed elements (LINEs). *See* Repetitive DNA
Long-terminal repeats (LTRs). *See* Repetitive DNA
Los Alamos National Laboratory (LANL) database, viral genome resources, 453–456
Lysosome, 677
Lysozyme, 592

Malaria parasites, eukaryotic genome sequencing, 572–574
Mammalian Gene Collection, gene expression bioinformatics, 168–169
Mann-Whitney tests, microarray data, 200
Manually curated database, multiple sequence alignment, 341
Mapping, human genome sequencing, genetic/physical distance mapping, 629
Mass spectrometry, high-throughput protein analysis, 251–254
Matrix-assisted laser desorption/ionization time-of-flight (MALDI/TOF) spectroscopy, high-throughput protein analysis, 249, 251–252
Matrix. *See* Basic Local Alignment Search Tool (BLAST)
  identity, 64
  multiplication, 53, 54, 56
  mutation probability matrix, 51, 53, 57–59
Maximum likelihood (ML), tree-building. *See* Molecular phylogeny
Maximum parsimony (MP). *See* Molecular phylogeny
MB, 27
Measles virus, 453–455, 457–458
Medical Subject Headings (MeSH), 35, 655–656
MEDLINE (Medical Literature, Analysis, and Retrieval System Online)
  PubMed, 24–25
  use, 35–36
Megablast algorithm, BLAST advanced searches, 136
Mendelian disorders, characteristics, 652
Messenger RNA (mRNA), *See* Ribonucleic acid, messenger. *See also* Gene expression.
Metacentric chromosome, defined, 565. *See also* Chromosome
MetaCyc database, cellular pathways, 259–262
MetaFam database, multiple sequence alignment, 339
Metaphase, chromosome analysis, 564. *See* Chromosome
Metazoans (animals), genome sequencing, 582–594
Methyl-CpG-binding protein domain (MBD), BLAST search, 227–229
Methyl-CpG-binding protein 2 (MeCP2)
  domains and motifs, 227–229
  gene silencing (transcriptional repression), 628
  Online Mendelian Inheritance in Man (OMIM) data, 659–664
  Rett syndrome, 657–667
Microarrays
  chromosomal abnormalities, 674
  chromosome 21 array, 557
  databases and repositories, 182, 184
  data analysis, 180–182
    annotation, 214–215
    ANOVA, 200
    artifacts, 191
    basic principles, 189–191
    Cluster/TreeView, 206, 208–209
    clustering, 189–190, 204–210
    curse of dimensionality, 203, 211
    Database Referencing of Array Genes Online (DRAGON), 214–215, 263
    descriptive (exploratory) statistics, 191, 203–214
    Euclidean distance, 203
    global normalization, 192–194
    hierarchical cluster analysis, 204–210
    inferential statistics, 191, 198–203
    k-means clustering, 210
    labeling efficiency, 191–192
    local normalization, 194–198
    log ratios, 192, 194
    Pearson correlation coefficient *r*, 204
    preprocessing, 191–198
    principal components analysis, 190, 211–213
    scatter plots, 193–198
    self-organizing maps, 210–211
    Significance Analysis of Microarrays (SAM), 200–203
    Standardization and Normalization of Microarray Data (SNOMAD), 178, 196–197
    supervised analysis, genes or samples, 213–214
    websites, 182, 216–217
  defined, 172–176
  experimental design, 177–178, 199–201
  biological confirmation, 182
  genomic, 674
  herpesviruses, 445–446, 450
  hybridization, 178–179
  image analysis, 179–180
  Microarray Gene Expression Data (MGED), 191
  vs. phylogenetic trees, 383
  protein microarray, 247
  replicates, 190
  RNA or probe preparation, 178
  tissue microarray, 247
  viral genomes, 458
  websites, microarray data repositories, 182, 184
Microbial genomes, sequencing history, 409–418

MicroRNAs (miRNAs), function, 553
Microsatellites. *See* Repetitive DNA
Microsporidia. *See Encephalitozoon cuniculi,* 529
Minimum Information About a Microarray Experiment (MIAME), gene expression bioinformatics, 182
Minisatellites. *See* Repetitive DNA
Mitochondrial DNA
  absent in some eukaryotic genomes, 570
  α-Protobacterium, eukaryotic genomes, 575
  completed genomes, 407–408, 410
  history, 407, 675
  human disease, 675–676
  human genome, 407–409
Molecular barcoding, genomic deletions, yeast strains, 526–528
Molecular clock hypothesis, molecular phylogeny and evolution, 360–363
Molecular Modeling Database (MMDB)
  National Center for Biotechnology Information (NCBI), 26
  Protein Data Bank (PDB) entries, 286, 290–291
Molecular phylogeny. *See* evolution
  analysis of bacteria/archaea, 478, 479
  analysis of eukaryotes, 504, 540
  analysis of fungi and *Encephalitozoon cuniculi,* 529
  analysis of plants, animals, and fungi, 575
  analysis of plants, 577
  analysis of primates, 595
  analysis of slime mold and metazoans, 582
  analysis of viruses, 439, 443, 451
  goals, 364–365
  historical background, 357–364
  insulin evolution in guinea pig and coypu, 359–360
  long-branch attraction, 384
  molecular clock hypothesis, 360–363
  multiple sequence alignment, 375–377
  neutral theory, 363–364
  operational taxonomic unit (OTU), 367–369
  programs
    GCG, 712–713
    Neighbor Treemaker, 383
    Phylogeny Analysis Using Parsimony (PAUP), 379–381, 384, 388
    PHYLIP, 377, 383
    TREE-PUZZLE, 386
  selection of DNA, RNA, or protein sequences, 371–374
  substitutions, synonymous and nonsynonymous, 371–374
  transitions and transversions, 373
  tree-building methods, 377–386
    character-based, 377–378, 383–386
    distance-based, 377–383
    Hamming distance, 378–379
    Maximum likelihood, 377, 386
    Maximum parsimony, 377, 383–386, 540
    Neighbor-joining, 377, 383
    UPGMA, 206, 377, 379–382
  tree evaluation, 386–389
    bootstrapping, 386–389
  trees
    bifurcating, structure and properties, 366–367
    clade, 367
    enumerating, 368–369
    gene/protein trees, structure and properties, 365–370
    inferred tree, 365
    nomenclature, 365–368
    outgroups, 368
    roots, 368
    species versus gene/protein, 369–370
    true tree, 365
  websites, 389
Molecule-specific Basic Local Alignment Search Tool (BLAST), advanced searching, 133, 169
Monera, 396, 399
Monogenic disorders. *See* Disease, monogenic disorders
Morbidity map, human mitochondrial genome, 675–676
Morbillivirus, 453
Mortality statistics, causes of death, 653
Motifs (fingerprints). *See* Protein domains/motifs
  analysis tools, 265
  defined, 225–226
  protein patterns, 231–233
Mouse (*Mus musculus*) genome
  C value paradox, 543
  genome sequencing, 589–591
  human disease models, 677, 679, 683–685
  Mouse Genome Informatics (MGI) website, 591
  Mouse Genome Sequencing Consortium
mRNA. *See* ribonucleic acid, messenger.
Mule, 565
Mutation probability matrix. *See* matrix
Multicellular organisms, genomic analysis, 401, 414, 582, 584
Multidomain proteins
  Basic Local Alignment Search Tool (BLAST) search parameters, 114–117
  description, 228–231
Multiple sequence alignment
  assessment, 346–350
  automated versus manual database curation, 341
  databases
    BLOCKS, 334
    Conserved Domain Database, 333, 339–340
    Domain Architecture Retrieval Tool (DART), 333, 340
    integrated sources (InterPro, MetaFam, iProClass), 336–340
    PopSet, 340
    Protein family (Pfam) database, 331–332
    PRINTS, 334
    PROSITE, 335
    Simple Modular Architecture Research Tool (SMART), 332
  defined, 320
  distance measures *vs.* similarity, 326
  Feng and Doolittle progressive sequence alignment, 321–325
  hidden Markov models, 325–329
  iterative algorithms, 345, 348
  phylogenetic analysis, 375–377
  programs, 345
    ClustalW and ClustalX, 343–344
    HMMER, 328–329
    PileUp (GCG), 345, 709–712
  typical uses, 321
  user-generated alignments, 341–349
MUMmer software
  Arabidopsis sequencing, 579
  prokaryotic genome analysis, 490–495
Munich Information Center for Protein Sequences (MIPS)
  Comprehensive Yeast Genome Database, 507–510, 512
  in GeneCards, 661
  plant genome sequencing, 578
Mus musculus. See Mouse
Mutation. *See* Molecular phylogeny
  database resources, 661–667
  defined, 648
  disease and DNA variation, 647–649
  Human Genome Project data, 609
  molecular phylogeny, 358, 361–362, 364, 372–375
Mutation probability matrix. *See* pairwise sequence alignment
Mycology, defined, 503

National Center for Biotechnology Information (NCBI), 10
  basic components, 24–26
    BLAST, 25
    data access example, 27–32, 33–35
    DNA/protein sequence data access, 27–35
    Entrez, 24, 31
    home page features, 24–26
    literature access, 35–36
    LocusLink, 27–29, 31
    OMIM, 25
    PubMed, 24, 35–36
    reference sequence (RefSeq), 29–30
    structure, 25, 286–294
    taxonomy, 26
    UniGene, 21–22, 31

UniGene and expressed sequence tags (ESTs), 20–22
UniGene and cDNA libraries, 164–168
human genome data, 609
Human Map Viewer, 609–613
human-mouse homology map, 590
pol (polymerase) gene example, 33–34
TaxPlot database, 489–490, 493
Trace Archive, 425–427
National Human Genome Research Institute (NHGRI)
Genome Hub, 608
HapMap project, 668
website, 609
National Institute of Allergy and Infectious Diseases (NIAID), viral classification, 438
National Institutes of Health Intramural Sequencing Center (NISC), eukaryotic genome comparisons, 567
National Library of Medicine (NLM), 35
MeSH terms, 655–656
PubMed, 24
Neanderthal DNA, 82
Needleman and Wunsch algorithm. *See* Global sequence alignment
Feng-Doolittle progressive sequence alignment, 321
pairwise sequence alignment, 63–68
Nematodes, genome sequencing, 583–585
Neurospora crassa, genomic analysis, 529–530
Neutral substitution rate, 589–591
Neutral theory of evolution, 363–364
Noncoding regions. *See* Repetitive DNA
gene definition, 551–553
human genome, 629–635
phylogenetic analysis, 371–374
Nonparametric testing
bootstrapping analysis, tree evaluation, 388–389
microarray data, 200
Northern blot, gene expression bioinformatics, 158–159, 164
N50 statistics, human genome sequencing, 625, 627
Nuclear magnetic resonance (NMR), protein structure, 282–285
Nucleic acids, biological databases for, 16
Null hypothesis testing
gene expression bioinformatics, cDNA libraries, 165–167
microarray data analysis, inferential statistics, 198–203
N value paradox, eukaryotic protein-coding genes, 562–563

O-notation, 71
Oak Ridge National Laboratory, human genome website, 11, 608, 642
Odorant-binding protein (OBP), 7, 38. *See* Lipocalins
Basic Local Alignment Search Tool (BLAST) query, 82, 117, 122
function, 245
gene, 547–549
structure, 274–275, 277, 282–283, 288–294, 299–302, 304–307
Olfactory receptors, 123, 268, 591
Online Mendelian Inheritance in Animals (OMIA), 677
Online Mendelian Inheritance in Man (OMIM)
description, 659–664
hemophilia data, 566
holdings, 664
links to Ensembl, 613
website at National Center for Biotechnology Information (NCBI), 25
Open reading frames (ORFs). *See* Gene
Operational taxonomic unit (OTU). *See* Molecular phylogeny
Operons, prokaryotic genome analysis, 482
Optimal alignment identification, pairwise sequence alignment, Needleman and Wunsch algorithm, 63, 67
ORF finder, protein-coding sequences, 558
Organelles
endoplasmic reticulum, 242, 281
eukaryotic genomes, 541
human disease pathways, 675–676
lysosome, 242
mitochondrion, 242, 675–676
nucleus, 543
peroxisome, 243, 570
Organism, bioinformatics and role of, 5–7
Organism-specific sites. *See* Basic Local Alignment Search Tool (BLAST)
Orphan sequences, multiple sequence alignments, 350
Orthologous sequences
defined, 43
Basic Local Alignment Search Tool (BLAST), 88
eukaryotic genome comparisons, 566–569
human disease genes, 676–685
pairwise sequence alignment, 43–50
prokaryotic genome analysis, 486–495
Oryza sativa. See Rice
Outgroup identification, phylogenetic tree roots, 368
Oxytocin, 360

Pairwise sequence alignment. *See* FASTA search algorithm, Multiple sequence alignment; PAM
BLAST 2 Sequences, 72–73
BLOSUM scoring matrices, 60–61
Dayhoff model
accepted point mutations, 50–60
evolutionary perspective, 365
detection limits, 61–62
dynamic programming, 63
protein vs. DNA alignment, 41–42
dot plots, 73, 75
evolution and homology, 47–50
gaps, 47
Genetics Computer Group (GCG) program, 44, 68–71, 701–709
homology, similarity, and identity definitions, 42–47
matrices
applications, 59–60
log-odds score matrix, 57–59
mutation probability matrix, 57–60
PAM250 and other, 53–60
Needleman and Wunsch global sequence alignment algorithm, 63–68, 321
percent identity, 73–77
protein structure, homology and functional genomics, 274–276
repetitive DNA sequences, 551, 554
Smith and Waterman local sequence alignment algorithm, 69–71
statistical significance testing, 73–77
global alignments, 76–77
local alignments, 77
scramble test, 76–77, 103
web resources, 78–79
PAM (accepted point mutation) matrices, 50–60
human disease models, 679, 683
pairwise sequence alignment, 50–60
applications, 59–60
Dayhoff model, 50–59
mutation probability matrix to log-odds score matrix, 57–59
Paralogous genes/proteins
defined, 43
pairwise sequence alignment, 43–47
Saccharomyces cerevisiae, 511–519
Paramecium, eukaryotic genome sequencing, 549, 574
Parametric testing, microarray data, 200
Parsimony analysis, see Maximum parsimony
Partitioning clustering techniques, microarray data analysis, *k*-means clustering, 210
Pathogenicity, bacteria, 466
Pathology, defined, 653
Pathophysiology, defined, 653
Pattern-hit initiated BLAST (PHI-BLAST), advanced searching, 145–147
PCA. *See* Principal components analysis
Pearson correlation coefficient, microarray data analysis, 204

Penetrance, complex genetic disorders, 671
Penicillin, 504
Percent identity, pairwise sequence alignment, 73–75
Percent similarity, pairwise sequence alignment, 46–47
Permutation test, microarray data analysis, inferential statistics, 200, 202–203
Peroxisome, 570
Pfam. *See* Protein family database
Phanerochaete chrysosporium (white rot fungus), genomic analysis, 530
Phenotype variations
  disease and, 647–649, 684–687
  human disease models, 676–679
PHI-BLAST. *See* Basic Local Alignment Search Tool (BLAST)
phosphoribosyl pyrophosphate, 261–262
Photolithography, microarray fabrication, 179
Phrap software
  genome sequencing, 422, 424
  human genome sequencing, 625
Phred software, genome sequencing, 422
PHYLIP program, molecular phylogeny, 377, 383
Phylogeny. *See* Molecular phylogeny
Phylogenetic trees. *See* Molecular phylogeny; tree of life
Phylogenomics, 429–430
Phylogeny Analysis Using Parsimony (PAUP) program. *See* Molecular phylogeny
PileUp (GCG) package, multiple sequence alignment, 345–347
  alternative algorithms, 345, 348
  phylogenetic analysis, 376
  user guidelines, 709–712
PipMaker, eukaryotic genome comparisons, 566–567
Plant genomes. *See also Arabidopsis*; Rice
  Arabidopsis thaliana (thale cress), 576–579
  chromosomes, sequencing history, 415–416
  overview and classification, 575–577
  Oryza sativa (rice), 545, 579–581
Plasmodium falciparum, eukaryotic genome sequencing, 572–574. *See vir*
Plastid
  defined, 573
  plant chloroplasts, 575–576
Ploidy, eukaryotic genomes, 543. *See* Chromosome; Diploid; Haploid
Point mutations, disease, 648
Pol (polymerase) gene/protein of human immunodeficiency virus
  pairwise alignment, 80–82
  properties, 7, 9
  protein domain/motifs, 229–232
  sequence access in databases, 33–35
  used as query sequence in Basic Local Alignment Search Tool (BLAST), 80–82, 114–117
Polyacrylamide gel electrophoresis (PAGE), high-throughput protein analysis, 248–252
Polymerase chain reaction (PCR).
  Electronic PCR sequence tag sites, 560
  genomic deletions, yeast strains, molecular barcoding, 526
  with reverse transcription (RT-PCR), 158
  yeast genetic footprinting, 522, 524
Polymorphism. *See* Single-nucleotide polymorphism (SNP)
Polyploidy, *Saccharomyces cerevisiae* duplication and gene acquisition, 512
P1-derived artificial clones (PACs), human genome sequencing, 621, 623
PopSet, multiple sequence alignment, 340, 342
Positionally cloned genes, human disease, 686
Position-specific iterated BLAST (PSI-BLAST)
  distantly-related protein searching, 137–144
    corruption errors, 144
    odorant-binding protein and retinol-binding protein, 275
    performance assessment, 143–144
  protein structure studies, 305, 310
Position-specific scoring matrix (PSSM)
  multiple sequence alignment, 233, 327
  position-specific iterated BLAST (PSI-BLAST), 97, 138–144
  reverse position-specific BLAST (RPS-BLAST), 333
Postranslational modifications, 226, 265–266
Prediction of Apicoplast Targeted Sequences (PATS), malaria genetics and genomics, 574
Pre-mRNA, mRNA gene expression studies, 160–161
Preprocessing techniques, microarray data analysis, 191–198
Prevalence, defined, 447
Primary protein structure. *See* Protein structure
Primates, 362, 401, 591–595
Principal components analysis (PCA)
  microarray data analysis, 190, 211–213
  Pfam database, 332, 337
  in analysis of viruses, 453
PRINTS database, multiple sequence alignment, 334
Prion proteins, 233, 312, 438
Probe preparation, gene expression studies. *See* Microarrays
Progenote, 398
Progressive sequence alignment, Feng-Doolittle technique, 321–325
Prokaryotic genomes. *See* Genomes, archaea and bacteria
Promoter regions, database resources, 563–564
PROSITE database
  multiple sequence alignment, 335, 341
  protein motifs, 231–233
Protein. *See* Cellular pathways; Gene Ontology; High-throughput protein analysis; Protein structure; Proteomics
  defined, 223
  domain and motifs, 225–233
    defined, 225–226
    Homo sapiens, 636, 639
    human proteome common domains, 227
    lipocalins, 231–233
    Saccharomyces cerevisiae, 506–508
    SMART database definitions, 226
  family
    defined, 225–229
    Homo sapiens, 636, 638
    InterPro database definitions, 226
    Saccharomyces cerevisiae, 506–508
  fingerprint, defined, 225
  fold, defined, 225
  function, 243–247
  localization, 242–243. *See* Transmembrane topology
  modular nature, 225–233
  physical properties, 233–237
  primary structure. *See* Protein structure
  profile, defined, 226, 233
  protein-coding gene, defined, 551–552
  protein-ligand interactions, web resources, 267
  protein-protein interactions, proteomics, 247–258
  repeat, defined, 225–226
  sequencing, Edman degradation, 250
  signature, defined, 225
  websites, 265–267
Protein-coding gene. *See* Annotation techniques; Gene; Protein
Protein Data Bank (PDB). *See* Protein structure.
Protein Family (Pfam) database
  description, 331–337
  Interpro constituent, 225
  molecular phylogeny, 365
  nematode genome sequencing, 584–585
Protein kinase C, 153, 232
Protein structure
  and disease, 311–312
  and homology and functional genomics, 274–276

Critical Assessment of Techniques for Protein Structure Prediction (CASP), 311
primary structure, 276, 278–279
Protein Data Bank (PDB), 26, 287–292
Access via NCBI, 289
FASTA search, 289
protein fold space, 292–293, 310–311
root-mean-square deviation (RMSD), 294, 306–309, 310
secondary structure, 276–281
structure databases
CATH database, 297–301
Dali Domain Dictionary, 301–304
FSSP database, 302, 305–307
Macromolecular Structure Database at European Bioinformatics Institute, 289
SCOP database, 293
SCOP used to assess PSI-BLAST, 143–144
VAST, 295
structure determination
ab initio prediction, 282, 309–310
comparative homology modeling, 282, 303–309
nuclear magnetic resonance (NMR), 282, 285
X-ray crystallography, 282–285
structure visualization
Chime visualization program, 292, 295
Cn3D, 291
interactive tools, 285, 291–292
SwissPDB viewer, 292, 294
target selection, 285–287
tertiary structure, 281
viral genomes, 458
website resources, 313
Proteomics
high-throughput protein analysis, 247–258
affinity chromatography and mass spectrometry, 252–254
false-positives and false-negatives results, 251–252, 253–254, 256
Rosetta stone approach, 256, 258
two-dimensional gel electrophoresis, 248–252
yeast two-hybrid system, 254–256
proteome, defined, 223–225
proteome, *Drosophila*, 586–587
proteome, human, 608, 636–641
proteome, malaria genetics and genomics, 573–574
proteome, *Saccharomyces cerevisiae*, 506–511
α-Protobacterium, eukaryotic genomes, 575
Protozoans, eukaryotic genome analysis, 568, 570–574
PRRP algorithm, multiple sequence alignments, 345
Pseudogenes
eukaryotic genomes, 547, 552
phylogenetic analysis, 373
PSI-BLAST. *See* Basic Local Alignment Search Tool (BLAST)
PSORT server, protein localization, 242–243
PSSM. *See* Position-specific scoring matrix
PubMed, 24, 35–36
P value, Basic Local Alignment Search Tool (BLAST), *E* values and, 106–107

Quagga, 351–352
Quantitative trait locus (QTL), complex genetic disorders, 669
Quaternary protein structure, 278
q value, Significance Analysis of Microarrays (SAM), 202

Rate of nucleotide substitution, molecular phylogeny, 360–364, 371–374
Rat genome, genomics resources, 594
Raw scores, Basic Local Alignment Search Tool (BLAST), 106
RBP. *See* Retinol-binding protein
RBP4. *See* Retinol-binding protein
Reference Sequence (NCBI RefSeq), database characteristics, 29–30
RefSeq. *See* Reference Sequence
RepBase Update, repetitive DNA sequencing, 550
Repeat-induced point mutation (RIP), *Neurospora crassa*, 530
RepeatMasker software, 550–556, 559, 561
Repetitive DNA
C value paradox, 543–544
DNA reassociation studies of Britten and Kohne, 544–545
eukaryotic genomes, 543–550
human genome, 620, 629–633
Human Genome Project conclusions, 608
interspersed repeats (transposon-derived repeats), 546–547, 630–632
Alu repeat sequences, 547, 553–554, 608, 631
DNA transposons, 547
long interspersed elements (LINEs), 547, 556, 631
long-terminal repeats (LTRs), 547, 556, 631
retrotransposon, defined, 545
Saccharomyces cerevisiae Ty elements, 510
short interspersed elements (SINEs), 547, 556, 631
structure, 631
transposons absent in *Giardia*, 571
transposons as molecular fossils, 631–632
processed pseudogenes, 547, 630, database, 547
segmental duplications, 547–549, 608, 632–633
Case Western Reserve University Duplication Browser, 547
simple sequence repeats, 547, 556, 632–633
microsatellites, 547
minisatellites, 547
triplet repeats, 547
tandem repeats, 549–550
α-satellite DNA, 550
telomeric repeats, 549
prokaryotic genomes, 492
Saccharomyces cerevisiae genome analysis, 509
software programs, 490–495, 550–556
Retinol-binding protein (RBP4; RBP)
access from databases, 27–33
Basic Local Alignment Search Tool (BLAST), 89, 108–113
description, 7
function analysis, 243, 245
GenBank database entries for, 19–21
high-throughput protein analysis, two-dimensional gel electrophoresis, 248–254
multiple sequence alignment, 49, 319, 321–326, 371–374, 375–376
gene, 558–563, 567
gene expression, 172
pairwise sequence alignment, 42, 46–48, 60, 68–71, 73–77
phylogenetic analysis, 379–389
position-specific iterated BLAST (PSI-BLAST), 138–144
PubMed search protocol, 35–36
structure and properties, 7, 241, 242, 245, 274, 293
homology and structural genomics, 274–276
Retrotransposon. *See* Repetitive DNA
RettBase, genetic information access, 666–667
Rett syndrome. *See* Disease, human
Reverse position-specific BLAST (RPS-BLAST), Conserved Domain Database (CDD), 333
Reverse transcriptase, 159. *See also Pol*
Reverse transcription polymerase chain reaction (RT-PCR). *See* Polymerase chain reaction
Rfam database, noncoding RNAs, 553
Rhodopsin, 112, 233
Ribonucleic acid (RNA)
human genome genes, noncoding RNAs, 634–635
Saccharomyces cerevisiae genome analysis, 507, 509

Ribonucleic acid, messenger (mRNA), 160–162. *See* Gene expression, Microarrays.
  correlation with protein expression, 174–176
  export, 161
  gene expression
    microarrays, 173–178
    preparation, 178
  processing, 160–161
  surveillance, 161
Ribonucleic acid, micro (miRNA), 553
Ribonucleic acid, ribosomal (rRNA)
  defined, 160
  Dictyostelium discoideum genome, 582
  European Small Subunit Ribosomal RNA Database, 401
  Fungi, 506–507, 509
  human genome genes, 634–635
  phylogenetic studies, 399, 569
  phylogenetic studies of bacteria and archaea, 474–475
  phylogenetic studies of animals and fungi, 575
  ribosomal gene clusters, tandem repeats, 549
  translation functions, 553
Ribonucleic acid, small nuclear RNAs (snRNAs), 553, 634–635
Ribonucleic acid, small nucleolar RNA (snoRNAs), 506, 553, 634–635
Ribonucleic acid, transfer (tRNA)
  defined, 160
  identification, 559
  human genome genes, noncoding RNAs, 635
  yeast, 506–507
Ribulose-1,5-bisphosphate carboxylase (rubisco), 541, 577
Rice (*Oryza sativa*)
  genomic analysis, 541, 579–581
  genome annotation, protein-coding genes, 560–562
Risk factors, complex genetic disorders, 670–672
RNA. *See* Ribonucleic acid
RNA interference, 585
RNA polymerase, 476
RNase protection assay, gene expression bioinformatics, 158
Root-mean-square deviation (RMSD), comparative modeling, 294, 306–309, 310
Root, phylogenetic trees, 368
Rosetta Stone technique
  high-throughput protein interaction analysis, 256, 258
  protein structure, 308–310
rRNA. *See* ribonucleic acid, ribosomal.
RTNBase, noncoding RNAs, 553
RT-PCR. *See* Polymerase chain reaction.
Saccharomyces cerevisiae (budding yeast)
  functional genomics, 520–528
    exogenous transposons, 523–526
    genetic footprinting with transposons, 522–523
    molecular barcoding, genomic deletions, 526–528
    multiprotein complexes, affinity chromatography, 252–254
    protein localization by epitope tagging, 243
    sporulation studied with microarrays, 193, 194
    yeast two-hybrid system, 254–256
  gene nomenclature, 514
  genome analysis, 505–519
    duplications, 511–519
    genome sequencing, 505–506
    genome features, 506–508
    typical chromosome, 508–511
  human disease models, 677–679
  web resources, 533
Saccharomyces Genome Database (SGD), 509, 513, 521, 522, 534
SAGA algorithm, multiple sequence alignments, 345
Satellite DNA. *See also* Microsatellites; Minisatellites
  eukaryotic genome sequencing, 550
Scatter plots. *See* Microarrays, data analysis
Schizosaccharomyces pombe (fission yeast)
  genomic analysis, 528, 530–531
  human disease models, 677, 681–682
  proteome size compared to *Plasmodium*, 573
  Saccharomyces cerevisiae comparison, 505
Scoring matrices. *See* Basic Local Alignment Search Tool (BLAST)
Scramble test, pairwise sequence alignment, 76–77, 103
Search parameters, Basic Local Alignment Search Tool (BLAST), 92–97
Search space, Basic Local Alignment Search Tool (BLAST), statistical significance tests, 105–106
Search strategies, Basic Local Alignment Search Tool (BLAST)
  general concepts, 108
  results evaluation, 108–113
Secondary protein structure. *See* Protein structure
Segmental duplication. *See* repetitive DNA
SEG program, repetitive DNA analysis, 94, 551
Selfish DNA. *See* Repetitive DNA sequences
Self-organizing maps (SOM), microarray data analysis, 210–211
Sensitivity, defined, 75–76
Sequence Search and Alignment by Hashing Algorithm (SSAHA), BLAST advanced searches, 136
Sequence-tagged sites (STSs)
  eukaryotic gene-finding, 560
  NCBI database, 22, 24
Serial analysis of gene expression (SAGE), 169–172
Sexual reproduction, eukaryotic genomes, 541
Shine-Dalgarno sequence, prokaryotic genome analysis, 480
Short interspersed elements (SINEs). *See* Repetitive DNA
Shotgun sequencing
  hierarchical shotgun sequencing, 421–423
    human genome sequencing, 620–626
  whole-chromosome method, 505, 573, 582
  whole-genome shotgun (WGS) sequencing
    Drosophila genome, 586
    genome sequencing strategy, 421–422, 505
    human genome, 620
Sickle cell anemia, 312, 314, 652, 656–657. *See* Disease, human
Signature, defined, 225. *See* Protein
Significance Analysis of Microarrays (SAM), basic principles, 200–203
Simian Immunodeficiency Virus. *See* Disease, animal
Similarity, pairwise sequence alignment, defined, 46
Simple Modular Architecture Research Tool (SMART) database
  multiple sequence alignment, 332–333, 338
  protein domain/motif definitions, 226
Simple sequence repeats. *See* Repetitive DNA
SIM program, pairwise sequence alignment, PAM matrix applications, 59
Sim4 program, BLAST advanced searches, 136
Single-gene disorders. *See* Disease, monogenic disorders
Single nucleotide polymorphism (SNP)
  human disease, 666–669
  in human genome, 609
  in mouse genome, 591
  in prokaryotes, 492
Slime mold. *See Dictyostelium discoideum*
Small nuclear RNAs (snRNAs). *See* Ribonucleic acid, small nuclear
Small nucleolar RNA (snoRNAs). *See* Ribonucleic acid, small nucleolar

Small-subunit ribosomal RNA. *See* Ribonucleic acid, small-subunit
Smith-Waterman algorithm, pairwise sequence alignment, 69–71
  relation to Basic Local Alignment Search Tool (BLAST), 87, 102
SNC1 and *SNC2*, yeast genes, 518–519
SNP. *See* Single nucleotide polymorphism
Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), high-throughput protein analysis, 248–249
Software packages. *See* specific programs; Websites
Sooty mangabeys, 447, 449
Speciation, 369–370
Species, defined, 442
Species trees, structure and properties, 369–370. *See also* Molecular phylogeny
Specificity, defined, 75–76
SSAP algorithm, protein structure, 297
SSO1 and *SSO2*, yeast syntaxin genes, 518–519
Standardization and Normalization of Microarray Data (SNOMAD), microarray gene expression studies, 178
  local normalization, 196–197
Stanford HIV RT database, 456, 460
Stanford Microarray Database, 182–183
  yeast gene expression, 194, 520
Statistical significance tests
  Basic Local Alignment Search Tool (BLAST)
    E values and, 103–106
    Raw and bit score analysis, 106–107
  microarray data analysis, inferential statistics, 198–203
  pairwise sequence alignment, 73–77
    global alignment, 76–77
    local alignment, 77
Structural genomics. *See* Protein structure
Structure databases. *See* Protein structure
Substitutions
  pairwise sequence alignment, 47
  phylogenetic analysis, 374
Superoxide dismutase *(SOD1)*, cellular pathways, 261, 264
Supervised data analysis, gene/sample classification. *See* Microarrays, data analysis
SwissPDB viewer. *See* Protein structure
SwissProt database
  Accession numbers, 27, 28, 33
  database entry, 230–231
  contents, 303
Synapomorphy, 372
Synaptosome-associated protein of 25 kilodaltons (SNAP-25), 233, 234, 236, 238
Syngamy, defined, 541
Syntaxin, 244, 268, 518–519
Synteny. *See* Conserved synteny
Systematics, historical background, 399

Tag sequencing, Serial Analysis of Gene Expression (SAGE), 169–170
TAIR. *See* The Arabidopsis Information Resource
Tandem repeats
  eukaryotic genome sequencing, 547–549
  Saccharomyces cerevisiae duplication and gene acquisition, 511, 514
Tandem Repeats Finder, eukaryotic genome sequencing, 547
Target selection/acquisition, three-dimensional protein structure. *See* Protein structure
Taxonomy, National Center for Biotechnology Information (NCBI), 26. *See* Genomes
TaxPlot
  prokaryotic genome analysis, 489–493
  Schizosaccharomyces pombe (fission yeast) data, 531–532
tblastn program. *See* Basic Local Alignment Search Tool (BLAST)
tblastx program. *See* Basic Local Alignment Search Tool (BLAST)
Tears, 147
Technical replicates, gene expression microarrays. *See* Microarrays, data analysis
Telomeres, eukaryotic genomes
  chromosome organization, 509, 549, 564
Tentative ortholog groups (TOGs), Basic Local Alignment Search Tool (BLAST), TIGR BLAST, 130–133
Tertiary protein structure. *See* Protein structure
Test statistics, microarray data. *See* Microarrays, data analysis
Tetrahymena, eukaryotic genome sequencing, 549, 574
Tetraploidy, *Saccharomyces cerevisiae* duplication and gene acquisition, 512
TFastA program, pairwise sequence alignment, 706–709
The Arabidopsis Information Resource (TAIR), 578–581
The Institute for Genomic Research (TIGR)
  archaeal genome sequencing, 413
  Aspergillus fumigatus genome data, 529
  Basic Local Alignment Search Tool (BLAST), 130–133
  Comprehensive Microbial Resource (CMR) database, 469–471, 475–476
  free-living organism genome sequencing, 409–413
  gene indices, 169
  genome sequencing resources, 421
  plant genome sequencing, 578
  Rice Genome Project, 580
Thioredoxin, 275, 279
Three-dimensional protein structure. *See* Protein structure
Threshold values, Basic Local Alignment Search Tool (BLAST), 101–103
Thymidine phosphorylase, lateral gene transfer, 483–486
TIGR. *See* Institute for Genomic Research (TIGR)
Time of divergence, molecular phylogeny, 364
TMPred, protein localization, 244
Topoisomerase, 256
Trait, defined, 655
Transcription. *See* Gene expression
Transcription factors, 563–564
Transcription factor binding sites, 321
Transfer RNA (tRNA). *See* Ribonucleic acid, transfer.
Transitions, phylogenetic analysis, 373–375, 379
Transmembrane topology
  prediction programs, 236–237, 244
  protein localization, 242
  web resources, 267
Transposons. *See* Repetitive DNA
  genetic footprinting, 522–523
  transposon tagging, 243, 523–526
Transversions, phylogenetic analysis, 373–375, 379
Tree-building. *See* Molecular phylogeny
TREE-PUZZLE program, maximum likelihood, 386
Tree of life, 6–7, 364, 365, 397, 399, 404, 431
Trees, phylogenetic. *See* Molecular phylogeny
Trigonencephaly, deletion 11q syndrome, 566
TRIPLES database, exogenous transposon harnessing, 525
tRNA. *See* Ribonucleic acid, transfer.
tRNAscan-SE search server, repetitive DNA sequencing, 552, 557
tRNA synthetase, 478
True tree, defined, 365
Trypanosomes, eukaryotic genome sequencing, 571–572
Tsetse flies, 571
t-test. *See* Microarrays, data analysis
Two-dimensional bacterial genomic display (2DBGD), prokaryotic genome analysis, 480
Two-dimensional gel electrophoresis. *See* Proteomics
Two-hybrid system. *See* Proteomics
Two-way hierarchical clustering. *See* Microarrays, data analysis

Unicellular pathogens, eukaryotic genomes, 571–574
UniGene project. *See* National Center for Biotechnology Information (NCBI)
Uniparental disomy, chromosome organization, 566
Universal Mutation Database, disease data, 666
University of California Santa Cruz Bioinformatics ("Golden Path"), human genome data, 609, 614–616, 622–623
Unweighted pair group method with arithmetic mean (UPGMA)
  Feng-Doolittle progressive sequence alignment, 323
  microarray data analysis, clustering techniques, 206
  tree-building, 377, 379–382
USAGE database, serial analysis of gene expression (SAGE), 172
User-generated multiple sequence alignment, 341–349

Vaccine-preventable diseases
  bacterial, 474
  viral, 441
VAMP/synaptobrevin protein, 518–519
  SNC1 and SNC2, *Saccharomyces cerevisiae,* 518–523
Vasopressin, 360
Vector, cloning, 163
Vector Alignment Search Tool (VAST). *See* Protein structure
Vesicles. *See also* syntaxin.
  Saccharomyces cerevisiae, 504, 518–519
  transport, 227
vir gene family, malaria genetics and genomics, 153, 574
Viral Genome Database, 456
Viral Genome Organizer, 455–456, 459
Viral genomes. *See* Genomes, viral.
Virion, basic properties, 437–438
Viroids, basic properties, 438
Virus. *See* Viral genomes
VISTA database, eukaryotic genome comparisons, 566–568
Vitamin A, retinol-binding protein (RBP), 7

Web exercises, 9–10
Websites
  Basic Local Alignment Search Tool (BLAST), 122–123
    pattern-hit initiated BLAST (PHI-BLAST), 146
    PSI-BLAST servers, 151
  BLOCKS, 60
  cancer, 689
  CATH database, 297–301
  cellular pathways, 267
  Censor, 550, 552
  chromosome databases, 565
  Cluster/TreeView, 206, 208–209
  Conserved Domain Database (CDD), 93, 330, 339–340
  Dali Domain Dictionary, 301–304
  Database Referencing of Array Genes Online (DRAGON), 214–215, 263
  dbSNP at NCBI, 669
  DNA Database of Japan (DDBJ), 10, 15, 37
  EcoCyc/MetaCyc, 269–262
  Entrez, 34
  ExPASy protein analysis, 11, 33, 59, 230–231, 235, 251–254, 257, 259, 292, 294
  European Bioinformatics Institute (EBI), 10, 15, 51, 508, 509
  FASTA, pairwise sequence alignment, 72
  FlyBase, 581
  GenBank database, 17, 37
  GeneCards, 661, 665
  Gene Ontology (GO) Consortium, 239
  Generic Model Organism Project, 581
  Genome Analysis Pipeline, Oak Ridge National Laboratory, 558–563
  GENSCAN, 557
  glossaries, 734
  GRAIL, 557
  GrailEXP, 559–563
  human disease, 675, 686
  Human Gene Expression Index (HuGE), 192
  Human Gene Mutation Database, 661
  human genome data, public access sites, 609, 642
  IMAGE, 169
  key bioinformatics sites, 10–11
  Kyoto Encyclopedia of Genes and Genomes (KEGG), 260, 263–264
  National Center for Biotechnology Information (NCBI), 10, 15
  Medical Subject Headings (MeSH) at National Library of Medicine, 35
  MEDLINE, 35
  Mouse Genome Informatics (MGI), 591, 592
  Munich Information Center for Protein Sequences (MIPS), 507–508, 512
  National Center for Biotechnology Information (NCBI), 10. *See* National Center for Biotechnology Information
  National Library of Medicine, 34
  Online Mendelian Inheritance in Man (OMIM), 659–664
  pairwise sequence alignment, 79
  phylogeny software, 389
  PipMaker, 567
  Protein Data Bank, 26
  PSI-BLAST, 138
  PubMed, 24, 37
  RepeatMasker, 550–556
  Saccharomyces Genome Database (SGD), 509, 513, 521, 522, 534, 581
  SCOP, 144
  Sequence Retrieval System (SRS) servers, 33
  serial analysis of gene expression (SAGE), 169, 172
  Significance Analysis of Microarrays (SAM), 200–203
  SIM, pairwise sequence alignment, 59
  Standardization and Normalization of Microarray Data (SNOMAD), 178, 196–197
  SwissModel, 309
  The Institute for Genomic Research (TIGR), 581
  TRIPLES, 525, 526
  tRNAscan-SE, 552, 557
  VISTA, 567, 568
  WormBase, 581, 584
Wellcome Trust Sanger Institute genome sequencing website, 419
  Saccharomyces pombe (fission yeast) data, 530
West Nile virus, 38, 587
Whole-genome duplication
  Arabidopsis thaliana, 578–579
  chromosome organization, 565
  Saccharomyces cerevisiae (budding yeast), 511–519
Whole-genome shotgun (WGS) sequencing. *See* Shotgun sequencing
Wilcoxon tests. *See* Microarrays, data analysis
World Health Organization (WHO)
  global burden of disease, 656–657
  ICD website, 657
  Leishmania genomics, 576
WormBase, nematode genome sequencing, 581, 584

X chromosome disorders, Rett syndrome, 657–659. *See* Disease, human
X-ray crystallography, protein structure, 42, 282–285

Yeast. *See* Fungi; Genomes, eukaryotic; specific yeasts, e.g. *Saccharomyces cerevisiae*

Z score, 76–77
Zinc finger domain, 227

Author Index

Adams, M. D., 164, 186
Altschul, S. F., 71–72, 83, 101–102, 105, 124
Anfinsen, C. (1916–1995), 40, 258

Baldi, P., 325
Baltimore, D., 608
Bateman, A., 331
Beadle, G., 529, 530
Bernardi, G., 428, 628
Berzelius, J. J. (1779–1848), 223
Birney, E., 325, 353, 609
Bishop, J. O., 164
Bork, P., 225, 269, 363
Botstein, D., 522
Brazma, A., 191
Brenner, S. E., 73–74, 83, 285
Britten, R. J., 544, 545
Brown, P. O., 173, 188, 208, 522
Büchen-Osmond, C., 438
Burley, S. K., 273, 315

Cantor, C., 358, 379
Casjens, S., 468, 497
Chakravarti, A., 608
Childs, B., 649, 691
Chothia, C., 73–74, 83, 144
Chou, P. Y., 279
Collins, F., 695
Cuvier, G. (1769–1832), 538

Darwin, C. (1809–1882), 357, 397
Dayhoff, M. (1925–1983), 50, 78, 83, 365
Da Vinci, L. (1452–1519), 2, 126
De Ketham, J. (active 1455–1470), 14, 188, 606, 646
Dickerson, R. E., 360
Dobzhansky, T. (1900–1975), 365
Doolittle, R. F., 62, 83, 225, 269, 321, 351, 353
Doolittle, W. F., 504

Eddy, S., 325, 328, 552
Eisen, J., 483
Eisen, M., 206, 208, 220

Fasman, G. D., 279
Felsenstein, J., 377, 389
Feng, D.-F., 321, 351, 353
Fitch, W., 43
Fodor, S., 186

Garrod, A. (1857–1936), 650–651
Gerstein, M., 315, 547
Gibson, T. J., 225
Golub, T., 210
Green, E., 608, 695
Gumbel, E., 86
Guttmacher, A., 695
Guyer, M., 695

Haeckel, E. (1834–1919), 358, 396
Haussler, D., 609, 614
Henikoff, J. G., 60–61, 83
Henikoff, S., 60–61, 83, 269
Higgins, D., 353

Jacq, B., 269
Jones, D. T., 315
Jukes, T. H., 358, 379
Jurka, J., 550

Karlin, S., 105
Karp, P. D., 258–260
Kellam, P., 461
Kendrew, J., 272, 283, 360
Kent, J., 136–137, 609
Kimura, M. (1924–1994), 358, 364
Kohne, D. E., 544, 545
Koonin, E. V., 273, 418
Krogh, A., 325
Kuchenmeister, F. (1821–1890), 502

Lancet, D., 661
Lander, E., 210, 424
Lederberg, J., 80, 530
Leeuwenhoek, A. van (1632–1723), 465, 505, 570
Linnaeus, C. (1707–1778),399, 401
Lipman, D., 71, 102, 124
Lupas, A., 236, 237

Maddison, D. R., 404
Makarov, V., 598
Mann, G. (1836–1916), 222
Marcotte, E. M., 258
McKusick, V. A., 659
Miller, O. W., 156
Miller, W., 136

Needleman, S. B., 63–67, 78, 83
Nuttall, G. (1862–1937), 356

Ohno, S. (1928–2000), 512
Orengo, C., 315
Owen, R. (1804–1892), 44, 465

Pauling, L. (1901–1994), 80, 360, 647, 657
Pearson, W. R., 71, 83, 124
Perutz, M., 272, 360
Poustka, A., 635

Roberts, R. J., 160

Sali, A., 315
Salzberg, S., 483, 491
Sanger, F., 40, 358
Schuler, G., 186, 609
Schwann, T. (1810–1882), 505
Searls, D. B., 598
Sharp, P. A., 160
Simpson, G. G. (1902–1984), 358
Smith, T. F., 69–71, 78
Snyder, M., 243
Speed, T., 191
Stein, L., 418
Strimmer, K., 386
Swofford, D., 379, 392

Tatum, E., 529, 530
Taylor, W., 353
Thompson, J. D., 343, 348
Thornton, J. M., 273, 315
Trent, J., 173, 188
Tuppy, H., 318

Valle, D., 649, 691
Velculescu, V., 186
Venter, J. C., 186, 608

Waterman, M. S., 69–71, 78, 424
White, O., 481
Woese, C., 399, 483
Wolfe, K. H., 515, 516
Wunsch, C. D., 63–67, 78, 83

Zuckerkandl, E., 80, 647