Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

IMMUNOINFORMATICS: BIOINFORMATIC STRATEGIES FOR BETTER UNDERSTANDING OF IMMUNE FUNCTION

The Novartis Foundation is an international scientific and educational charity (UK Registered Charity No. 313574). Known until September 1997 as the Ciba Foundation, it was established in 1947 by the CIBA company of Basle, which merged with Sandoz in 1996, to form Novartis. The Foundation operates independently in London under English trust law. It was formally opened on 22 June 1949.

The Foundation promotes the study and general knowledge of science and in particular encourages international co-operation in scientific research. To this end, it organizes internationally acclaimed meetings (typically eight symposia and allied open meetings and 15–20 discussion meetings each year) and publishes eight books per year featuring the presented papers and discussions from the symposia. Although primarily an operational rather than a grant-making foundation, it awards bursaries to young scientists to attend the symposia and afterwards work with one of the other participants.

The Foundation's headquarters at 41 Portland Place, London W1B 1BN, provide library facilities, open to graduates in science and allied disciplines. Media relations are fostered by regular press conferences and by articles prepared by the Foundation's Science Writer in Residence. The Foundation offers accommodation and meeting facilities to visiting scientists and their societies.

Information on all Foundation activities can be found at
http://www.novartisfound.org.uk

Novartis Foundation Symposium 254

IMMUNOINFORMATICS: BIOINFORMATIC STRATEGIES FOR BETTER UNDERSTANDING OF IMMUNE FUNCTION

2003

John Wiley & Sons, Ltd

Published in 2003 by John Wiley & Sons Ltd,
The Atrium, Southern Gate,
Chichester PO19 8SQ, UK

National 01243 779777
International (+44) 1243 779777
e-mail (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on http://www.wileyeurope.com
or http://www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Novartis Foundation Symposium 254
viii+263 pages, 32 figures, 11 tables

Library of Congress Cataloging-in-Publication Data

Immunoformatics : bioinformatic strategies for better understanding of immune function / [editors, Gregory Bock and Jamie Goode].
p. cm. – (Novartis Foundation symposium ; 254)
Includes bibliographical references and index.
ISBN 0-470-85356-5 (alk. paper)
1. Immunoinformatics. I. Bock, Gregory. II. Goode, Jamie, III. Series.
QR182.2.I46I46 2003
571.9′6–dc22 2003057599

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 470 85356 5

Typeset in 10½ on 12½ pt Garamond by Dobbie Typesetting Limited, Tavistock, Devon.
Printed and bound in Great Britain by T. J. International Ltd, Padstow, Cornwall.
This book is printed on acid-free paper responsibly manufactured from sustainable forestry, in which at least two trees are planted for each one used for paper production.

Contents

Symposium on Immunoinformatics: bioinformatic strategies for better understanding of immune function, held at the Novartis Foundation, London, 8–10 October 2002

Editors: Gregory Bock (Organizer) and Jamie Goode

This symposium is based on a proposal made by Nikolai Petrovsky and Vladimir Brusic

Participants

Stephan Beck Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Massimo Bernaschi IAC 'Mauro Picone' (C.N.R.), Viale del Policlinico 137, I-00161 Rome, Italy

Francisco Borras-Cuesta Department of Internal Medicine, School of Medicine, University of Navarra, Irunlarrea 1, 31008 Pamplona, Spain

Vladimir Brusic Knowledge Discovery Department, Institute for Infocomm Research, 21 Heng Mui Keng Terrace, 119613, Singapore

Annie De Groot Brown University, TB/HIV Research Laboratory, Brown University, Box G, Providence, RI 02912 , USA

Charles DeLisi Center for Advanced Genomic Technology, Boston University, 1st Floor, Room 102, 48 Cummington Street, Boston, MA 02215, USA

Darren R. Flower Bioinformatics Group, The Edward Jenner Institute for Vaccine Research, Compton, Newbury, Berkshire RG20 7NN, UK

Kamalakar Gulukota gvk bioSciences Private Limited, #210, 'My Home Tycoon', 6-3-1192, Begumpet, Hyderabad 500 016, India

Paul Kellam Virus Genomics and Bioinformatics Group, Department of Immunology & Molecular Pathology and Department of Virology, Windeyer Institute of Medical Sciences, Windeyer Building, 46 Cleveland Street, London W1T 4JF, UK

Can Kesmir Department of Theoretical Biology, Utrecht University, Padualaan 8, 3584 CH, Utrecht, Netherlands

Marie-Paule Lefranc IMGT, the international ImMunoGeneTics information system®, Université Montpellier II, Laboratoire d'ImmunoGénétique

Moléculaire, LIGM, UPR CNRS 1142, Institut de Génétique Humaine, 141 rue de la Cardonille, F-34396 Montpellier Cedex 5, France

Tim Littlejohn Biolateral, PO Box A51, Enfield South, NSW, 2133, Australia

Terry Lybrand Department of Chemistry, Vanderbilt University, Center for Structural Biology, 5142 Biosci/MRB III, Nashville, TN 37232-8725, USA

Hanah Margalit Department of Molecular Genetics & Biotechnology, Hebrew University Hadassah Medical School, PO Box 12272, Ein Kerem, Jerusalem 91120, Israel

Steven G. E. Marsh Anthony Nolan Research Institute, Royal Free Hospital, Pond Street, Hampstead, London NW3 2QG, UK

Alan S. Perelson MS K710, T-10, Theoretical Division, Los Alamos National Laboratory, PO Box 1663, Los Alamos, NM 87545, USA

Nikolai Petrovsky Canberra Hospital, Autoimmunity Research Unit, PO Box 11, 2606 Woden, ACT, Australia

Hans-Georg Rammensee (*Chair*) Interfakultäres Institut für Zellbiologie, Abteilung Immunologie, Universität Tübingen, Auf der Morgenstelle 15, D-72076 Tübingen, Germany

Lukas Roth Novartis Pharma, Transplantation Research, WSJ-386.9.26, CH-4002 Basel, Switzerland

Diego Silva (*Novartis Foundation Bursar*) Autoimmunity Research Unit, The Canberra Hospital, Canberra, ACT 2065, Australia

Christian Schönbach RIKEN Genomic Sciences Center, Biomedical Knowledge Discovery Team, E-209, 1-7-22 Suehiro-cho, Tsurumi, Yokohama, Kanagawa, 230-0045, Japan

Stefan Stevanović Interfakultäres Institut für Zellbiologie, Abteilung Immunologie, Universität Tübingen, Auf der Morgenstelle 15, D-72076 Tübingen, Germany

Edgar Wingender GBF-Braunschweig, Genome Analysis, Mascheroder Weg 1, D-38124 Braunschweig, Germany

Dominik Wodarz Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, MP-655, Seattle, WA 98109-1024, USA

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

Chair's introduction

Hans-Georg Rammensee

Interfakultäres Institut für Zellbiologie, Abteilung Immunologie, Universität Tübingen, Auf der Morgenstelle 15, D-72076 Tübingen, Germany

This is a timely meeting. Although Vladimir Brusic's opening paper is titled 'Immunoinformatics — the new kid in town', this is actually a field that has been around for a while, although under a different name. At least part of what we know of as immunoinformatics was previously known as 'theoretical immunology'. There was an important meeting on this subject in New Mexico in 1988, which resulted in a two-volume book (Perelson 1988).

The subject of immunoinformatics as we see it today can roughly be divided into three areas: the hard, the soft and the semi-soft. A challenge for this group is to decide by the end of the meeting whether I am correct with this classification! Let me start with a description of hard immunoinformatics. This contains what I will call 'hard facts': DNA, RNA and peptide sequences that we can write down. This part of immunoinformatics can be used for a growing number of applications that will have a direct impact on biomedicine. One example is peptides for T cell recognition, working out which peptides are recognized by the T cell receptor during an infection. Hard immunoinformatics is one of the newest parts of the field and is only a few years old. The amount of information in this realm is growing exponentially. 15 years ago all we had were a few DNA sequences, but now we have a tremendous amount of data stored in various databases.

Semi-soft immunoinformatics comprises algorithms and parameters which we use to create the 'hard' part. It includes all the prediction algorithms we use in DNA or peptide sequences: we say that a particular DNA sequence will interact with some regulatory protein or this piece of protein sequence will interact with the MHC. The one hallmark of this semi-soft area is that all the predictions can be tested accurately. You can predict the peptide sequence to bind to HLA, and then go on and test whether this is true. Some of the predictions will be correct and others won't. At one point, though, we may get to a stage where we can omit the verification of the prediction by experiment. I personally think this will never be the case, and we will always have to verify our predictions, but others may disagree.

Then we come to the soft part of immunoinformatics. This is I would to define as something that can never be tested with hard facts. This may raise some

controversy. I would classify this part of immunoinformatics as what has previously been known as 'theoretical immunology'. This includes mathematical descriptions of the behaviour of populations, whether this is at the level of the individual, or at cellular or antibody levels. It involves interactions between antibodies, infectious agents and T cells. I would like to propose that these kinds of models will stay soft because it is not possible to verify the predictions experimentally. If you predict that you need 30 T cells in a human to start an efficient immune response against a viral infection using mathematical modelling, you will never be able to prove this. On the other hand, while these predictions cannot be tested accurately, they can certainly be of help. For example, if one can calculate in a mathematical model the percentage of people that need to be immunized against measles to avoid an epidemic, this will be of great use.

So I propose that it is useful to break down immunoinformatics into these three categories of hard, semi-soft and soft. At the end of the meeting we can discuss whether or not my proposal is correct. Two important questions related to this are whether soft immunoinformatics can ever be tested accurately, and whether the predictions from semi-soft immunoinformatics can stand alone without experimental verification. Let's now move to the first presentation.

Reference

Perelson AS (ed) 1988 Theoretical immunology. Proceedings of the Theoretical Immunology Workshop, June 1987, Santa Fe, New Mexico. Addison-Wesley, Reading, MA

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

Immunoinformatics — the new kid in town

Vladimir Brusic*† and Nikolai Petrovsky†‡

**Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, †Centre for Medical Informatics, Division of Science and Design, University of Canberra, Bruce ACT 2617 and ‡National Health Sciences Centre, Canberra Clinical School, Woden ACT 2606, Australia*

Abstract. The astounding diversity of immune system components (e.g. immunoglobulins, lymphocyte receptors, or cytokines) together with the complexity of the regulatory pathways and network-type interactions makes immunology a combinatorial science. Currently available data represent only a tiny fraction of possible situations and data continues to accrue at an exponential rate. Computational analysis has therefore become an essential element of immunology research with a main role of immunoinformatics being the management and analysis of immunological data. More advanced analyses of the immune system using computational models typically involve conversion of an immunological question to a computational problem, followed by solving of the computational problem and translation of these results into biologically meaningful answers. Major immunoinformatics developments include immunological databases, sequence analysis, structure modelling, mathematical modelling of the immune system, simulation of laboratory experiments, statistical support for immunological experimentation and immunogenomics. In this paper we describe the status and challenges within these sub-fields. We foresee the emergence of immunomics not only as a collective endeavour by researchers to decipher the sequences of T cell receptors, immunoglobulins, and other immune receptors, but also to functionally annotate the capacity of the immune system to interact with the whole array of self and non-self entities, including genome-to-genome interactions.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 3–22

Biotechnology has provided methods and instrumentation for analysis and manipulation of biological systems on a massive scale. Information technology has provided hardware and software that enable data processing at an unprecedented speed and efficiency. Bioinformatics, defined as the storage, manipulation and interpretation of biological data (MacLean & Miles 1999), has emerged at the interface of life and information sciences. Bioinformatics has evolved as a crucial methodology in genomics, proteomics, and structural

biology. Immunoinformatics (also known as computational immunology) is a subset of bioinformatics focusing on the field of immunology. Immunoinformatics applications are increasingly becoming important to immunological research. The major findings of structural, functional and regulatory aspects of molecular immunology, coupled with the rapid accumulation of immunological data have been complemented by the development of more sophisticated computational solutions for immunology research.

Immunology is essentially a combinatorial science. The diversity in the human immune system is enormous — the total number of combinatorial arrangements of immunoglobulins (Ig) in an individual is greater than 10^9 (Jerne 1993). The T cell receptor (TCR) diversity in humans has been estimated (Arstila et al 1999) at between 10^7 and 10^{15} different clonotypes. There are approximately 10^{12} B cell clonotypes in an individual human (Jerne 1993). More than 500 allelic variants of class I human histocompatibility complex (MHC) molecules characterized to date allow theoretically more than 10^{13} class I haplotypes. The theoretical number of linear epitopes composed of nine amino acids, common targets in cellular immunity, is of the order 10^{11}. The number of conformational epitopes is far higher. These crude numbers, reflecting the complexity of the immune system in a very simplistic manner, indicate its enormous diversity. This diversity underpins our ability to discriminate between friend (self) and foe (non-self) and mount appropriate immune responses. Additional information includes multi-step processing pathways, network-type interactions, complex signalling and mechanisms for modulation of immune responses. Currently available data represent only a tiny fraction of possible situations and the amount of information will keep growing. With the steadily increasing amount of immunological information our ability to decipher the specific mechanisms of immune responses or correct undesirable immune responses is increasingly dependent on exploiting immunoinformatics strategies.

A major role of immunoinformatics is the management and analysis of immunological data with the basic infrastructure comprising numerous immunology database systems (Brusic et al 2000). Immunology databases provide access to, data extraction from, and analysis of immunological data. Standard bioinformatics methods, e.g. sequence analysis (Foster & Chanock 2000) and structural methods, e.g. structure modelling (immunoglobulin, Martin et al 1989; MHC, Schueler-Furman et al 1998, Rognan et al 1999; or TCR, Garcia et al 1998) are routinely applied to immunology studies. More advanced analyses of the immune system using computational models typically involve conversion of an immunological problem to a computational one, solving the computational problem, and translating the results into biologically meaningful interpretations. Examples include data-driven modelling of peptide binding to MHC molecules (Brusic et al 2001), theoretical modelling and complex analysis of the immune

system (Perelson 1989, Kepler & Perelson 1993), and statistical support for immunological experimentation (Merrill 1998). Virtually every aspect of immunology research uses some form of immunoinformatics. The appropriate use of informatics techniques has potential, as supported by examples of practical applications, to vastly improve the efficiency of immunology research. Complete genomes of more than 900 viruses and more than 80 microbes have been sequenced to date (Wheeler et al 2002). High-throughput approaches such as microarray technology (Glynne & Watson 2001), proteomics (Marshall & Williams 2002) and large-scale T cell epitope screening (Schönbach et al 2002) provide for genomic-scale screening and study of the immune system, and its role in beneficial and pathological immune responses. Practical immunoinformatics applications include screening of genomes for vaccine components (De Groot et al 2002), disease-specific gene expression (Saito 2001), studies of cell differentiation pathways, tolerance/immunity decision process and B cell transformation (Glynne & Watson 2001), antibody recognition site identification (Yoshimori & Del Carpio 2001), and integration of data into high level models of the immune system (Yates et al 2001). In the following sections we describe the status and challenges within the subfields of immunoinformatics and discuss the prospects for future developments.

Immunoinformatics

The immune system is intertwined with all other body systems. Bioinformatics applications are relatively well developed for some immunological areas, such as databases (Brusic et al 2000), genomic applications (Glynne & Watson 2001), study of T cell epitopes (Brusic & Zeleznikow 1999), or modelling immune responses (Bernaschi & Castiglione 2002). In other fields of immunology bioinformatics applications are still in their infancy, such as analysis of allergenicity of proteins (Gendel 2002) or proteomics (Klade 2002). Because of the combinatorial nature of immunological data, the importance of efficient, accurate and comprehensive use of immunoinformatic tools will continue to grow in importance for support of immunology research.

Immunological databases

Both molecular biology and immunology produce large amounts of data that have to be stored in general-purpose and specialist immunological databases. General-purpose biological databases contain annotated entries of biological sequences. These entries typically contain the sequence, a short description, the source organism, a list of structural or functional features and literature references. The major public databases include the nucleotide or protein

sequence databases GenBank/GenPept (*www.ncbi.nlm.nih.gov/Genbank/index.html*), EMBL/TrEMBL (*www.ebi.ac.uk/embl*), DDBJ/DAD (*www.ddbj.nig.ac.jp*), PIR (*www-nbrf.georgetown.edu*), SWISS-PROT (*www.expasy.ch/sprot*), PDB (*www.rcsb.org/pdb*), PROSITE (*www.expasy.ch/prosite*) and KEGG (*www.genome.ad.jp/kegg/kegg2.html*). The nucleotide databases — Genbank, EMBL and DDBJ — focus on collecting, annotating, and providing access to the entries of DNA sequences and the related information. GenPept, TrEMBL and DAD are protein databases derived from the translations of coding sequences of the three main nucleotides databases. SWISS-PROT and PIR are protein databases that are manually annotated. Their content is of higher quality than GenPept, TrEMBL and DAD, but they contain fewer entries. PDB is a database of 3D molecular structures. The PROSITE database contains biologically significant patterns and motifs. The KEGG databases comprise repositories on molecular interaction networks, chemical compounds and reactions relevant to cellular processes, and genomics data.

General-purpose databases contain large numbers of immunologically relevant entries and are invaluable resources, therefore, for immunology research. They do not, however, provide sufficient detail on immunological function. Specialist immunology databases provide more detailed information on immunologically relevant molecules, systems and processes. They are typically annotated by experts and contain immunology-specific annotations. Kabat database (*kabatdatabase.com*) contains entries of proteins of immunological interest: Ig, T cell receptors (TCR), major histocompatibility complex (MHC) molecules and other immunological proteins. The IMGT databases (*imgt.cines.fr*) contain high-quality annotations of DNA and protein sequences of Ig, TCR and MHC. They also contain IMGT-related genomic and structural data. The FIMM database (*sdmc.lit.org.sg/fimm*) focuses on protein antigens, MHC molecules and structures, MHC-associated peptides and relevant disease associations. The SYFPEITHI database (*syfpeithi.bmi-heidelberg.com*) contains entries of MHC ligands and peptide motifs. The HIV molecular immunology database (*hiv-web.lanl.gov/immunology*) is an annotated searchable repository of HIV1 T cell and B cell epitopes. More detailed reviews of important immunological databases and related issues can be found in (Brusic et al 2000, Petrovsky & Brusic 2002). The important database issues relate to data standardisation, data quality, interpretation of database entries, and the quality of computational tools for data extraction and analysis (Petrovsky & Brusic 2002), which will be discussed later in this text.

Bioinformatics applications to the study of T cell epitopes

The identification of T cell epitopes relies heavily on bioinformatics for initial screening followed by experimental validation. MHC molecules bind short peptides produced mainly by intracellular (MHC class I) and extracellular (MHC

class II) degradation of proteins and display them on the cell surface for recognition by the T cells (using TCRs) of the immune system. Binding of peptides to the MHC molecule is a prerequisite for immune recognition, but the number of peptides that can bind to a specific MHC molecule is limited. Peptides that bind specific MHC molecules are involved in initiation and regulation of immune responses. Determining peptides that bind specific MHC molecules is important for understanding immunity and has applications to vaccine discovery and design of immunotherapies. The combinatorial nature of this problem makes computational approaches necessary for systematic mapping of T cell epitopes.

Prediction methods are based on binding motifs (Rammensee et al 1999), quantitative matrices (Parker et al 1994) or higher complexity prediction models such as artificial neural networks (ANN) (Brusic et al 2001), hidden Markov models (HMM) (Brusic et al 2002) or molecular modelling (Schueler-Furman et al 1998, Rognan et al 1999). The binding motif describes amino acids commonly occurring at particular positions within peptides that bind to a specific MHC molecule. Quantitative matrices provide coefficients for each amino acid and each position within the peptide that can be used with appropriate formulae to calculate scores that predict peptide binding. The artificial intelligence methods of ANNs and HMMs are based on higher order models that can capture non-linear dependencies in the data sets. The data-driven models (binding motifs, quantitative matrices, ANNs and HMMs) are derived from experimental data sets and can be used for large-scale screening of potential vaccine components (Schönbach et al 2002, De Groot et al 2002). The important property of these models is that each binding motif can be encoded as a quantitative matrix, and each quantitative matrix can, in turn, be encoded as an ANN or a HMM. The accuracy of data-driven methods depends on the complexity of the model relative to the complexity of the peptide–MHC interaction, and on the quantity and representativeness of the data available for building a particular model. Molecular modelling methods utilise comparative modelling where known crystal structures and protein-peptide interactions are used as templates for building 3D models of molecular structures. If initial structural data are not available, *ab initio* modelling based on atomic simulations and residue statistics can be used. Molecular modelling is useful for detailed analysis of specific 3D structures and interactions, but being computationally intensive it is less useful for large-scale screening. Molecular modelling can be used for building complex data-driven methods, such as those for prediction of promiscuous MHC-binding peptides (Brusic et al 2002), or quantitative structure–activity relationships (QSAR) for vaccine discovery (Doytchinova & Flower 2002). The main issues for prediction of MHC-binding peptides are the quality, quantity, and representativeness of data available for model development, the complexity of

the selected predictive model relative to the natural complexity of the peptide–MHC interaction and the training and testing of the predictive model.

Mathematical modelling of the immune system

Observations of immune responses and cellular interactions at the organism level produce definite measurements, but are difficult to interpret at the molecular level. An example is the idiotypic network theory (Jerne 1993) which can be translated into speculative explanations at the molecular level. Mathematical modelling implemented as computational programs can easily translate speculative hypotheses into quantitative descriptions (Perelson 1989). The parameters of the mathematical models can easily be tuned to represent real behaviour of the immune system. These models can then be used for determining the framework for study of the kinetics of immune responses and practical applications such as prediction of immune interventions. Mathematical models of the immune system can model interactions of a large number of elements (10^6 or higher) thereby approaching the complexity of the human immune system. Remarkably accurate simulations using mathematical models have been developed for study of B cell (Kepler & Perelson 1993) and T cell responses (Coussens & Nobis 2002). More specific examples (Yates et al 2001) include modelling of tumour necrosis factor oscillations in allografts, differentiation of T helper cells (Th1/2), modelling T cell memory and cross-talk between TCRs.

Systemic level mathematical models provide a framework for understanding of the immune system as whole. We foresee the convergence of mathematical models at the systemic and molecular level in the future. Huge experimental data sets produced by genomics, proteomics and molecular biology efforts will ultimately be integrated with mathematical models of the immune system at the organism level to produce models of whole organism.

Emerging applications of immunoinformatics

Genomics focuses on the study and characterization of the complete set of DNA sequences (genome) from an organism. Similarly, proteomics focuses on study and characterization of the full protein complement of the genome. Following successful integration of bioinformatics in various fields of molecular biology, notably genomics and proteomics, immunoinformatics is the next frontier, namely the integration of bioinformatics with immunology. A major function of the immune system is to help the organism maintain homeostasis while interacting with self and foreign entities. Beneficial immune responses are targeted towards maintaining homeostasis, while pathological immune responses result in disease states, such as allergies or autoimmunity. The emerging field of immunomics

encompasses the genomics and proteomics of the immune system (Glynne & Watson 2001, Marshall & Williams 2002, Saito et al 2001, Coussens & Nobis 2002, Zagursky & Russell 2001). Immunomics focuses not only on deciphering the sequences of immunoglobulins and various cellular receptors, but is also instrumental for functional annotation of the immune system interactions with the whole array of self and foreign entities, including complete genome-to-genome interactions. Examples of fields that are expected to show rapid growth are immunoinformatics of disease (allergies, cancer, autoimmunity, infectious diseases), host–pathogen interactions, animal immunology, improved predictions of organ rejections, cytokine signalling and other regulatory network analysis, among others. In respect of development of immunoinformatics tools, we expect to see the integration of immunological databases with generic interfaces and ultimately the integration of system level mathematical models with molecular level models leading to applications in the development of novel therapeutic regimens and disease management.

Unifying concepts

The main issues that need to be resolved are those of common data standards, data quality and the accuracy of computational methods. These issues are critical for establishing a common immunoinformatics platform and enabling efficient and adequate use of immunoinformatics resources.

Standardization

Biochemical and molecular biology terms have been standardized by nomenclature committees, such as IUPAC/IUBMB (*www.chem.qmw.ac.uk/iubmb/nomenclature*). The gene ontology consortium (*www.geneontology.org*) has produced a dynamic controlled vocabulary of genes and proteins that can be applied to all organisms in rapidly changing environments. The immunogenetics ontologies and nomenclature for immunoglobulins have been defined recently (Ruiz & Lefranc 2002) and are accessible at the IMGT database. The HLA nomenclature system has been well-defined and accepted (*www.anthonynolan.org.uk/HIG/nomen/nomen_index.html*). Although the MHC nomenclature for other organisms has been under development (e.g. swine and bovine leukocyte antigens) a unifying system for the MHC nomenclature is lacking. Cytokine and cytokine related gene nomenclature is also not well defined — a comprehensive list of cytokine names can be found at the COPE web site (*www.copewithcytokines.de*).

In addition, each immunological database has its own unique structure, data models, and interfaces. Common interfaces, such as SRS (*srs6.ebi.ac.uk*) can integrate multiple databases and search tools, but are general tools. A common

interface for multiple immunological tools and databases would provide long-term benefits for immunology research. This common interface would provide seamless access to data and easy integration of both general and specialist bioinformatics tools.

Data issues

The interpretation of data extracted from the databases is highly dependent on the skills and knowledge of the user. In many cases the complicating factors are lack of standards, ad hoc nomenclature, variable quality annotations of database entries, incomplete data and biases embedded in the data. The optimal database searching tools for addressing a particular problem may require careful selection as well as setting of search parameters. Although the situation is slowly improving, the lack of bioinformatics education represents a serious obstacle to extracting the best value from data and unfortunately this problem often goes unnoticed by users. Data residing in databases are not of uniform quality, and even well-curated databases contain numerous errors (for a case study of errors in databases, see Srinivasan et al 2002).

Accuracy of computational methods

Hundreds of bioinformatics tools are available for analysing biological data. Many of these, such as sequence comparison and sequence alignment tools (such as standard bioinformatics tools BLAST or FASTA) calculate the distance between the query sequence and the database entries. This distance is based on user-selected parameters of the search and statistical assessment of the data and method. Therefore, search results may differ and assessing the accuracy of these tools is not informative. On the other hand, assessment of accuracy of predictive methods (such as prediction of peptide binding to MHC molecules) is of critical importance. In the past, most of the predictive models were generated and provided to the research community without careful assessment of their predictive performance. This resulted in some predictions of poor accuracy and a low level of acceptance of predictive bioinformatics models by the majority of researchers. More recently, assessment of predictive performance has become standard and vastly improved and refined predictive methods are appearing. A comparative study of the predictive performance of various methods has been recently published (Yu et al 2002). In addition, it was shown that predictive methods, when combined with experimental research in a cyclical fashion (Fig. 1.) can significantly improve the efficiency of research (Brusic et al 2001).

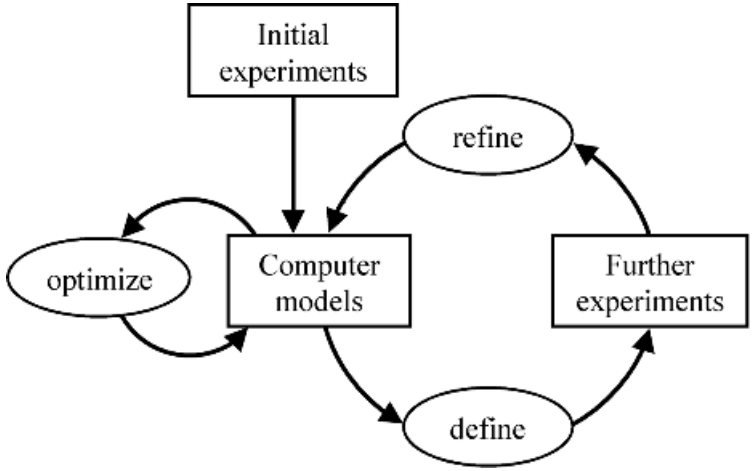


FIG. 1. Cyclical refinement of computer models used to define further experiments, including the optimization step. The optimization uses computer science methods, while refinement uses new experimental data.

Conclusion

Immunoinformatics is an enabling technology that will increasingly dominate immunology research, following the pattern set by genomics and proteomics. The scope of immunoinformatics is huge — it comprises databases, molecular-level and organism-level models, genomics and proteomics of the immune system, as well as genome-to-genome studies. Immunomics is thus the natural extension of genomics and proteomics and includes the study of organism-to-self and organism-to-organism interactions.

The efficient development and use of immunoinformatics will require the coordinated efforts of immunologists and bioinformaticians to establish common standards and protocols as well as standardized tools and interfaces. While coordinating efforts may be a challenge in this fast developing field, it is essential if we are to make sense out of the mountains of immunological data that will be produced in coming decades.

References

Arstila TP, Casrouge A, Baron V, Even J, Kanellopoulos J, Kourilsky P 1999 A direct estimate of the human $\alpha\beta$ T cell receptor diversity. Science 286:958–961

Bernaschi M, Castiglione F 2002 Selection of escape mutants from immune recognition during HIV infection. Immunol Cell Biol 80:307–313

Brusic V, Zeleznikow J 1999 Computational binding assays of antigenic peptides. Lett Pept Sci 6:313–324

Brusic V, Zeleznikow J, Petrovsky N 2000 Molecular immunology databases and data repositories. J Immunol Methods 238:17–28

Brusic V, Bucci K, Schönbach C, Petrovsky N, Zeleznikow J, Kazura JW 2001 Efficient discovery of immune response targets by cyclical refinement of QSAR models of peptide binding. J Mol Graph Model 19:405–411, 467

Brusic V, Petrovsky N, Zhang G, Bajic VB 2002 Prediction of promiscuous peptides that bind HLA class I molecules. Immunol Cell Biol 80:280–285

Coussens PM, Nobis W 2002 Bioinformatics and high throughput approach to create genomic resources for the study of bovine immunobiology. Vet Immunol Immunopathol 86:229–244

De Groot AS, Sbai H, Aubin CS, McMurry J, Martin W 2002 Immuno-informatics: mining genomes for vaccine components. Immunol Cell Biol 80:255–269

Doytchinova IA, Flower DR 2002 Quantitative approaches to computational vaccinology. Immunol Cell Biol 80:270–279

Foster CB, Chanock SJ 2000 Mining variations in genes of innate and phagocytic immunity: current status and future prospects. Curr Opin Hematol 7:9–15

Garcia KC, Degano M, Pease LR et al 1998 Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen. Science 279:1166–1172

Gendel SM 2002 Sequence analysis for assessing potential allergenicity. Ann NY Acad Sci 964:87–98

Glynne RJ, Watson SR 2001 The immune system and gene expression microarrays—new answers to old questions. J Pathol 195:20–30

Jerne NK 1993 The Nobel Lectures in Immunology. The Nobel Prize for Physiology or Medicine, 1984. The generative grammar of the immune system. Scand J Immunol 38:1–9

Kepler TB, Perelson AS 1993 Cyclic re-entry of germinal center B cells and the efficiency of affinity maturation. Immunol Today 14:412–415

Klade CS 2002 Proteomics approaches towards antigen discovery and vaccine development. Curr Opin Mol Ther 4:216–223

MacLean M, Miles C 1999 Swift action needed to close the skills gap in bioinformatics. Nature 401:10

Marshall T, Williams KM 2002 Proteomics and its impact upon biomedical science. Br J Biomed Sci 59:47–64

Martin AC, Cheetham JC, Rees AR 1989 Modeling antibody hypervariable loops: a combined algorithm. Proc Natl Acad Sci USA 86:9268–9272

Merrill SJ 1998 Computational models in immunological methods: an historical review. J Immunol Methods 216:69–92

Parker KC, Bednarek MA, Coligan JE 1994 Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J Immunol 152:163–175

Perelson AS 1989 Immune network theory. Immunol Rev 110:5–36

Petrovsky N, Brusic V 2002 Computational immunology: the coming of age. Immunol Cell Biol 80:248–254

Rammensee HG, Bachmann J, Emmerich NP, Bachor OA, Stevanović S 1999 SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50:213–219

Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V 1999 Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. J Med Chem 42:4650–4658

Ruiz M, Lefranc MP 2002 IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. Immunogenetics 53:857–883

Saito H, Nakajima T, Matsumoto K 2001 Human mast cell transcriptome project. Int Arch Allergy Immunol 125:1–8

Schönbach C, Kun Y, Brusic V 2002 Large-scale computational identification of HIV T-cell epitopes. Immunol Cell Biol 80:300–306
Schueler-Furman O, Elber R, Margalit H 1998 Knowledge-based structure prediction of MHC class I bound peptides: a study of 23 complexes. Fold Des 3:549–564
Srinivasan KN, Gopalakrishnakone P, Tan PT et al 2002 SCORPION, a molecular database of scorpion toxins. Toxicon 40:23–31
Wheeler DL, Church DM, Lash AE et al 2002 Database resources of the National Center for Biotechnology Information: 2002 update. Nucleic Acids Res 30:13–16
Yates A, Chan CCW, Callard RE, George AJT, Stark J 2001 An approach to modelling in immunology. Brief Bioinform 2:245–257
Yoshimori A, Del Carpio CA 2001 Automatic epitope recognition in proteins oriented to the system for macromolecular interaction assessment MIAX. Genome Inform Ser Workshop Genome Inform 12:113–122
Yu K, Petrovsky N, Schönbach C, Koh JYL, Brusic V 2002 Methods for prediction of peptide binding to MHC molecules: a comparative study. Mol Med 8:137–148
Zagursky RJ, Russell D 2001 Bioinformatics: use in bacterial vaccine discovery. Biotechniques 31:636–640

DISCUSSION

Petrovsky: I would like to start in a slightly argumentative mode, by questioning the idea that Hans Georg Rammensee brought up in this introduction that peptide binding data constitute hard evidence and immunoinformatic predictions constitute semi-soft or soft evidence. I would argue that the quality of data is dependent on the level of its validation rather than whether it is derived from laboratory studies or computer models. Hence, couldn't well validated computer algorithms be considered hard and poorly validated experimental assays be soft?

Rammensee: It is a matter of quality control.

Petrovsky: Exactly. The quality of the data is a reflection of their statistical validation rather than their source. As an example, consider how MHC restriction was originally described: when did this evidence go from being soft to being hard? We initially started with Zinkernagel and Doherty's original description of MHC restriction of viruses, but the nature of the molecules involved and the manner in which they interacted was pure conjecture. Over time experimental details led to the proposal that a complex of MHC, antigen and a TCR underpinned this phenomenon. At that stage, however, given that no-one had actually seen an MHC molecule or a TCR, was this hard or soft evidence of the existence of these molecules. Later there was argument about how MHC was binding antigens with some people believing the peptide was bound in the cleft whereas others thought it was bound outside the cleft. More recently crystal structures have begun to appear and for the first time we can actually visualize what, up to that point, people had been hypothesizing about. Hence substance is a question of validation. Sometimes we fool ourselves into thinking that because something was measured in a lab it must be hard, whereas

if it is derived from a computer model it must be soft. I do not think this reasoning is correct.

Gulukota: I would add that when you have an interface between computational biology and bench biology, often the computational side believes that 10 computations are not as good as a single experiment, but this could just be because they don't know that much about experiments. On the other side, however many experiments the experimentalists do, they don't quite believe it until a computer prediction says something similar. There needs to be a cultural shift. Hard and soft is very much in the eye of the beholder. When we talk about biology, it is pretty much all soft!

Rammensee: I was restricting the use of 'hard' to just DNA, RNA and peptides sequences. The hard facts about MHC restriction are the sequence of the MHC, the sequence of the peptide and the sequence of the TCR.

DeLisi: In effect, you are distinguishing data from concepts.

Gulukota: Even in data there are gradations of softness. If you consider data such as MHC peptide binding, there are three or four different ways of measuring this. I'm sure we all have our personal preferences about whether IC_{50} is better than half-life, for example. Until we have a consensus, we can't even call experimental data hard.

Brusic: I have experience with assessing which method is best for measuring peptide binding. I started from the computational end and interviewed people who measure peptide binding and asked them which method they considered the best. I got a unified answer, 'mine'! Then I took a fuzzy approach to interpreting measurement data by converting all the values to approximate measures of values.

Rammensee: I don't think the peptide binding is the most important component of the quality of a certain peptide. The most important part is whether this peptide is recognized under physiological conditions. If you have a virus-infected cell and a T cell, does the T cell that is specific for a particular peptide recognize the virus-infected cell? This is the acid test. We again come to the point about what the right test is and what the best criteria are for calling something solid or soft.

Stevanović: It is still difficult to judge the properties of peptides. You say that sequences are hard data, and I agree with this. But the properties of peptides in terms of binding to MHC molecules or recognition by TCRs vary with the experimental setting. In particular, in cancer immunology, we know very well that there are so-called T cell epitopes that do not function in many labs. Even T cell recognition can't be called 'hard' data.

Rammensee: We are talking about immunoinformatics, but sequences in databases are hard data.

Littlejohn: I think your concept of hard data is a useful one. Hard data should be seen as discrete information, observations that can be digitized, and that are qualitative and not continuously variable. 'Hard data' are the foundation stones

in molecular observations. Then, on top of hard data, we can superimpose 'noise' and biological variation, and the contextually dependent observations that we have discussed here. If this is what you mean by 'hard immunoinformatics data' then I think this is an extremely useful concept that constitutes a good reference point against which we can compute (i.e. carry out rigorous immunoinformatics).

DeLisi: Of course, there is noise in the hard data also. Sequencing has an error rate of about one base in a thousand.

Rammensee: This brings us back to the issue of quality control.

Marsh: I like this idea of hard data. The HLA database that we run is a 'hard database': it is a database of sequences. The difficulty we have is knowing how to link our hard database to other databases. For example, there are many databases doing peptide prediction for MHC binding peptides. Which one should we link up with? We don't want to link our hard database with a semi-soft database that gives poor predictions.

DeLisi: There needs to be more benchmarking. We have done this with peptide MHC. Zhiping Weng and her colleagues have an algorithm that is about 90% reliable in terms of both specificity and sensitivity. This has been benchmarked in terms of all the standard algorithms on the web. Parker comes close to that. If we have more benchmarking like this, then this will go some way to alleviating this problem.

Littlejohn: I think the problem is elsewhere: it lies with evidence. Many of the databases do not ascribe evidence as to how the information was derived. Was it experimental? What experiment? Was it computational? What method was used? Who did it, when, and in what context? This is the big problem. The Gene Ontology consortium is battling with this issue of 'evidence', and this consortium has only just begun to think about how to ascribe evidence codes to the methods used to assign function to genes. I'd argue that this is one of the great problems in bioinformatics in general, and it needs to be tracked in the database as well as the derived information.

Wingender: That was exactly the point I was going to make. When we start differentiating between hard, soft and semi-soft data, we have to assign where the 'facts' come from. What is the source of the experimental or computational evidence? Whenever we model these data and provide them through a database, we simply have to provide the evidence, along with these data and facts. We then need to try to make a quality assignment to the data on the basis of this information. I would like to add a caution here against databases that have been made using data collected in an automated manner. There are some terrible mistakes in these. The data must be extracted manually from the literature, but this is also an error-prone process. The original data in the paper can even contain errors. At some point we have to rely on the quality control step of peer-review, though.

Rammensee: With regard to the problem of interconnectivity of databases, I would say that if I had a database which is quality controlled and contains good data, I don't want to have it connected with a bad database — for instance, one made automatically without adequate curation. I would like to protect my database from being corrupted by poor data. Thus we need to discuss the two important issues of interconnectivity and quality control together.

Margalit: We all agree about the need for quality control and good documentation. Who can do this? Most of the databases are assembled by research groups and are not commercial. I know from other fields that I am involved in, such as transcription factor binding and protein–protein interactions, that these databases may start in the academy, but at some point they decide they can't handle the scale of the database and they make a consortium or go commercial. Perhaps this meeting represents an opportunity to think how we can best develop a single, quality controlled immunoinformatic database that isn't spoiled by bad data.

Borras-Cuesta: I have a point about the quality control of databases. One issue is whether a peptide binds or does not bind to MHC: one should control this. The other thing, related to the predictive algorithms, is how these peptides were defined and collected. You could have a database that tells you the truth with respect to binding, but which is skewed with respect to predicting the set of potential binders. This is very important. People who like us work in the induction of immune responses, and have to try to characterize a peptide to induce a response, go through all the steps predicting this with one algorithm and then another. By the end we do not trust any in particular. We use several algorithms, and select the peptides predicted with higher scores from all algorithms. These peptides are synthesized and tested in binding assays, if these are available, or used in immunization experiments. But if algorithms are going to be described which are potent, one should discuss how to build a good database. That is, a database which has no bias for a particular set of peptides because it has been built up using, ideally, several methods (i.e. peptides eluted from MHC molecules, identified with phage display libraries, using peptide libraries, etc.). Peptides from this database could then be used to develop an algorithm for the prediction of binding to MHC molecules.

Rammensee: You raise the important point that predictions can be tested.

Borras-Cuesta: Yes, we predict and then we test in a binding assay. This is not enough, of course, because they could be cryptic peptides. But if we predict and then it binds, then we use it.

DeLisi: The assay has to be quality controlled also. For instance, take the assay used by Parker. He validates it, but when you look at it you find this validity holds only under a certain range of conditions on the rate constants. Something may look valid, you do an analysis of it, and you find there is only some domain of validity.

The first thing that needs to be done, therefore, is to benchmark the assays. Then you benchmark the algorithms on benchmarked assays.

De Groot: I would like to second the idea of having a collective database. I would suggest that we categorize the peptides in the database by peptides that bind MHC and by peptides that are recognized by T cells. I agree that the type of assay is very important in this respect. We all train or benchmark our algorithms on different sets of peptides. We are now finding that the set of epitopes versus the set of binders might be slightly different subsets of HLA binding peptides. I have a second comment: I also wanted to mention that on Vladimir Brusic's time-line, the date that the structure of HLA was published by Don Wiley should be highlighted. When I first met Hannah Margalit and Charles DeLisi in Jay Berzovsky's laboratory, we were talking about how the peptide bound to the groove, and we were discussing about the peptide not being aligned with the sides of the groove. Once the crystal structure was published, this showed everyone the fact that the peptide was aligned parallel to the side of the HLA, and was also tightly constrained within the groove. This was a turning point for the field.

DeLisi: There was no doubt that the peptide was linear; the question was how it was oriented.

Rammensee: In the 1987 crystal structure (Bjorkman et al 1987), it was not clear how the peptide was organized.

Borras-Cuesta: This raises the point of how the peptide is read by the MHC II. In principle, it is possible that the peptide could be read from C-terminus to N-terminus in some circumstances. This is relevant to predictions. Someone should do the following experiment. Synthesize for instance 20 peptides known to be recognised by a given MHC II molecule. These peptides should also be synthesized in the C-terminal to N-terminal sense (that is, with the same amino acid sequence, but read from the C-terminus to the N-terminus, and not in the conventional way N-terminus to C-terminus). If some peptides from this new set were immunogenic in the context of the same MHC II molecule, then predictions should also take into account peptide sequences read from C-terminus to N-terminus.

De Groot: One thing we should add to the databases is information about non-binding peptides. We are all constrained by finances and we don't make the peptides that we predict wouldn't bind, because it is expensive to make them. However, many of us have done assays and found that some of the peptides that we have predicted don't bind. Some of us also make 'negative control' peptides and test these. It will be important to include the negative sets in the databases in order to improve the accuracy of our epitope prediction tools.

Rammensee: The quality of data will be worse if you include non-binding peptides, because the peptide-binding assay might miss some non-binders. What we call 'non-binding' peptides might bind if the assay conditions are altered.

Kellam: What we have been discussing are quality issues. Anyone who has been following the microarray field for the last few years will have seen how people have gone to extreme in describing how to 'quality control' experiments, to the point where you try to document absolutely everything. There is a huge community effort to describe standards and common protocols. In the end, if people start documenting what they are doing experimentally you have a chance of getting to the context of the data in the databases. For example, how many people even know the sex of the cell lines that they are working with? This can become important.

Littlejohn: I would like to comment on that from a standards and sociology point of view. The micoarray MIAME standard is supposed to be a minimum standard, yet it is often referred to by the user-community as a 'maximum irritation' standard, as it requires the biologists to capture more information than they ordinarily might. With regard to the database integration issue, eight years ago I attended the 'Meeting for the Interconnection of Molecular Biology Databases' (see *http://megasun.bch.umontreal.ca/ogmp/abstracts/mimdb.html* on the 'Organelle Genome Megasequencing Program') where many of these issues were discussed. There are a couple of developments in molecular biology databases that would be useful for us to consider by the immunoinformatics community. First, back then Peter Karp proposed that bioinformatics research would benefit from having a unified system of data interchange standards. However, as this idea was discussed, it became clear that each database curator has their own set of objectives, and so was unlikely to redesign their systems to fit a broad-community-developed standard that did not meet their narrower goals. The concept of bioinformatics databank warehouses has been around for a long time and has not made much headway into the community, primarily due to the fact that most databanks have evolved in isolation and have their own schema and specific target audiences, making their absorption in to a warehouse problematic. In spite of this, there has been a vast amount of effort put into systems that allow databank interconnectivity, such as the SRS system (Etzold et al 1996). Databank integration does not come at a quality cost. For example, SWISS-PROT, EMBL and GenBank all have databank cross references and these simply allow cross-databank navigation. Databank integration and 'ontological normalization' (deriving a common set of key-terms for accessing information across databanks) is a vigorous area of research, with many technologies variously called 'wrappers' or 'agents' serving as 'middleware' (software that joins other pieces of software or data) in this area. Interconnectivity is a critical issue, and it isn't in and of itself a problem. The final comment I have is that the debate should continue to focus on biology and not technology, although as Vladimir Brusic points out, at the end of the day this is a technology, a means to an end. Immunoinformatics is all about technologies that underpin the study of immunology, immunology is the

rationale, and immunoinformatics provides a means to probe problems in immunology.

Gulukota: I would like to second the comment you made about the microarray field. This is a much younger field than immunology, yet last month they had a report on an XML standard called MAGE-ML (Spellman et al 2002; *http://genomebiology.com/2002/3/9/research/0046.1*) for describing the hybridization protocol, the data and the databases for microarrays. If we ever produce an 'immunobank' where we deposit this data, we will need some standard such as this that would give us interoperability and connection, without at the same time contaminating good data with bad.

DeLisi: There are parts of this that are worth emulating, and in particular the sociological components. But there are some problems with the scientific parts. You could do the standardization as much as you want, but this assay has 90% false positives. We don't want to get into a situation where we are documenting errors. This is why I raised the issue of benchmarking the assays: those assays have not been benchmarked.

Gulukota: However, you cannot even begin benchmarking until everyone documents the data.

Borras-Cuesta: The quality of the assays is crucial. I'll give an example. We started with a peptide which is restricted to HLA-A2. We measured binding of this peptide to HLA-A2 using the T2 cell assay, but could not estimate its IC_{50} of binding because the peptide does not bind well to HLA-A2. However, in spite of its poor binding, the peptide is immunogenic. This means that peptides that can be negative in this assay can still be immunogenic. One mustn't forget that the immune response involves not only binding to MHC, but also recognition by the TCR. A peptide which binds not so well, but which is well recognized by the TCR, will be immunogenic. This is also a relevant point for predictions. One should consider the amino acids that point to the TCR and try to make predictions according to that.

Petrovsky: I think what everyone is saying here is consistent. Perhaps we should classify immunoinformatics processes into a hierarchy of confidence levels. Thus predictions for example of MHC–peptide binding, if the method is robust and well validated could be accepted at a high level of confidence approaching that of experimentally measured binding affinities. Other less well validated predictions would be afforded a lower level of confidence. A good example is the FDA system for regulating drugs. The FDA has defined minimum acceptable standards whether it be for the laboratory (good laboratory practices, GLP), manufacturing (good manufacturing practices, GMP) or clinical trials (good clinical trial practices, GCP). This is what Vladimir Brusic was alluding to when he talked about good prediction practices (GPP). For instance, if you want to get a drug registered with the FDA, everything to do with the

development and manufacture of this drug has to be done under good practices, i.e. GLP, GCP and GMP. By building up a hierarchy of well defined standards you ensure that the product you end up with can be trusted all the way back down through the system. This is what we need to be able to do with immunoinformatic tools if we are truly to be able to trust their outputs.

Littlejohn: High quality data was one of the gold standards of the Human Genome Project. However, there is a real danger with preventing release of data before it is considered 'perfect'. If over stringent quality control had been imposed at every step of the data generation cycle, the genomics community would never have had access to data such as the high-throughput Genome Survey Sequences portion of GenBank (*http://www.ncbi.nlm.nih.gov/dbGSS/index.html*). The research community demanded genomic data within 24 h of it being produced by the sequencing machines in the genome centres. The community understood the quality was lower, and treated it accordingly. It is the assignment of quality to the data that is important here. While we need 'gold standard' high quality data, there is also a need for rapid publication of lower quality data as long as it is assigned as such.

Brusic: These are sequence data; we are also discussing functional issues here.

Littlejohn: Yes, but surely there must be lots of non-sequence data that the immunoinformatics community wants to throw out in a quick and dirty fashion, so that researchers can have rapid access to this information?

Petrovsky: I disagree. If someone publishes a bad paper it can damage that whole area of endeavour. There are many examples of this. Poor quality T cell suppression papers in the 1980s damaged the area so badly that suppressor T cells were a dirty word for the next 20 years and very few researchers were brave enough to persist in the area. Twenty years on, we have slowly rediscovered suppressor T cells, but this would have all happened much more quickly if it hadn't been for the original poor quality data that damaged the area.

Littlejohn: If this was the case, then I would argue that these data were assigned too high a level of quality. If the data had been assigned a low-quality rating then it might still have been useful to someone.

Petrovsky: The difficulty is deciding who is to assign the quality ranking. You have to develop systems that validate the quality as the data are generated and the results of this ranking need to be widely available. Many of the suppressor cell data that were eventually discredited were published in very high ranking journals so even peer review of publications may be insufficient to ensure high quality.

Gulukota: It doesn't have to be personal, as in one person deciding whether data are good or bad. You ask the author to describe their methodology,

which automatically ascribes the level of trust someone wants to put in it. With high-throughput sequencing, everyone knows what it is and you can believe it or not depending on your comfort level.

Rammensee: To answer Nikolai Petrovsky's concern, we don't have the problem of bad papers being published that destroy the field. At the moment, no one can publish a paper saying that they predict a certain peptide will bind somewhere without any experimental data proving this prediction. The paper will not be accepted.

Petrovsky: There are currently no agreed statistical standards for handling predictions. If there isn't an agreed statistical criterion, the reviewers might think that *t*-tests are fine for assessing the accuracy of predictions and not understand relative operating characteristic (ROC) analysis which should be the gold standard for assessing predictions.

DeLisi: There is less of a problem with publication than there is with the web. On the web, anyone can put up a database. This is where the standard has to be. Perhaps there needs to be an indication of whether or not the data have been reviewed, and by whom.

Gulukota: We can't prevent anyone putting up databases, but at least if there is one database that all the community knows is documented in a certain way, people would go there preferentially.

Perelson: I believe that publishing predictions with or without experimental data depends somewhat on the group. There are modelling groups that make predictions but who do not have experimental collaborators. There may be some value in having them publish their predictions so that they are available to the community, and other people can then test them. The hard part is getting the theory and experiment connected, so people read both the predictions and the experimental validations.

Borras-Cuesta: If the databases are based on experimental facts, there is not much danger of a peptide being a binder or not. The problem there is we need to know which are the non-binders for the predictions. A panel of non-binder peptides is necessary because these peptides are needed to test the specificity of the algorithm. After all, we wish to accurately predict binder but also non-binder peptides.

Kesmir: This idea of good prediction practice sounds good, but it is not enough just to stipulate that a certain statistical test needs to be used or specific performance measures should be given: it depends solely on the test set that you are using. We have been discussing MHC predictions. We are also doing some proteasome predictions where the data are extremely sparse. We have to test our methods on just two or three peptides. I wonder whether what we are discussing is a realistic approach. It needs to be done, but we need to realise we are still working in a data sparse area.

References

Bjorkman P J, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC 1987 Structure of the human class I histocompatibility antigen, HLA-A2. Nature 329:506–512

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

Spellman PT, Miller M, Stewart J et al 2002 Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 3:RESEARCH0046

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

The future for computational modelling and prediction systems in clinical immunology

Nikolai Petrovsky*, Diego Silva*† and Vladimir Brusic‡

*Medical Informatics Centre, University of Canberra, Bruce ACT 2601, *Autoimmunity Research Unit, The Canberra Hospital, Woden ACT 2606, †John Curtin School of Medical Research, Canberra ACT 2606, Australia and ‡Institute for Infocomm Research, 21 Heng Mui Keng Terrace, 119613 Singapore*

Abstract. Advances in computational science, despite their enormous potential, have been surprisingly slow to impact on clinical practice. This paper examines the potential of bioinformatics to advance clinical immunology across a number of key examples including the use of computational immunology to improve renal transplantation outcomes, identify novel genes involved in immunological disorders, decipher the relationship between antigen presentation pathways and human disease, and predict allergenicity. These examples demonstrate the enormous potential for immunoinformatics to advance clinical and experimental immunology. The acceptance of immunoinformatic techniques by clinical and research immunologists will need robust standards of data quality, system integrity and properly validated immunoinformatic systems. Such validation, at a minimum, will require appropriately designed clinical studies conducted according to Good Clinical Practice standards. This strategy will enable immunoinformatics to achieve its full potential to advance and shape clinical immunology in the future.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 23–42

The explosive growth in biotechnology combined with major advances in information technology is producing vast quantities of readily accessible biological data with direct relevance to immunology research and clinical practice. New data are being added at an exponential rate through initiatives such as the Human Genome Project, Mouse Genome Initiative and Functional Annotation of the Mouse (FANTOM2) Project, amongst others. These initiatives focus predominantly on biological sequences, i.e. biological structures, genetic and physical maps, and pathways. Bioinformatic resources include immunological databases and computational methods for data extraction and analysis. Bioinformatic tools provide a means for fast and comprehensive

extraction of biological sequence information for comparison, analysis or interpretation. They also assist with the planning and design of laboratory experiments and thereby have the potential to accelerate knowledge discovery. The ability to efficiently extract and analyse useful information from the rapidly expanding number of databases is crucial for immunology research and ultimately for immunology clinical practice.

The size of the human genome is between 30 000 and 100 000 genes. A major current effort is to identify all these genes, and elucidate the structure and function of the proteins they encode. Many of these genes will be involved in immune function and a subset of these in immune-related diseases. Bioinformatic tools for database searching and biological sequence analysis allow quick identification of sequences of interest and provide substantial bibliographic, taxonomic or feature information. Tools for sequence comparison, motif searching, or profiling assist researchers in identifying biologically relevant sequence similarities as well as a new generation of bioinformatic tools that enables modelling of biological interactions and simulation of laboratory experiments.

Difficulties in the application of computational tools arise from the fact that most immunology researchers and particularly clinicians have only a limited understanding of sophisticated data analysis and their applicability and limitations, whilst most computer scientists lack understanding of the depth and complexity of immunological data (Petrovsky & Brusic 2002). Therefore, success in applying bioinformatics to immunology relies heavily upon individuals and groups who are able to cross the divide between these two disparate fields. The focus of this paper is the potential ability of immunoinformatics to transform clinical immunology practice and research. This point is illustrated with a number of key examples including the use of computational immunology and database mining to predict renal transplantation outcomes, identify novel genes potentially involved in immunological disorders, better understand disease relationships in HLA antigen presentation pathways and predict allergenicity.

Clinical practice whilst reliant upon research for advances is however guided by a different set of principles such that research proof may not always translate into clinical acceptance. The reasons for this are many and varied but in part relate to the central tenet of clinical practice which is 'first do no harm'. Also, clinicians by their nature tend to be conservative and skeptical of 'miracle' cures or advances. Hence, it will not be easy to get immunologists to accept the results of black boxes such as artificial neural networks or other computational prediction or modelling systems. This is particularly true if we wish them to accept computer based predictions or guidance in their clinical decision-making process. The only way forward therefore is to develop appropriate immunoinformatics frameworks and clinical standards capable of satisfying even the most skeptical clinician. Immunoinformatics

methods need to be validated to minimum clinical as well as experimental standards. This would require, for example, that all clinical modelling or prediction systems should be validated where possible by blinded, prospective clinical studies performed according to Good Clinical Practice standards. By applying equally rigorous scientific and clinical standards to immunoinformatics practices we will ensure widespread acceptance of this new field by scientists and clinicians alike.

Immunoinformatics in improving transplantation outcomes

Renal failure is an increasing problem around the world, with a rising incidence largely due to the rising incidence of type 2 diabetes. Although dialysis is a short-term solution, renal transplantation remains the optimum solution both for restoring quality of life and for increasing life expectancy of patients. A major limitation to renal transplantation is the supply of donor kidneys. Although success rates from renal transplantation continue to improve, a significant number of donor kidneys continue to be lost due to rejection or recurrent disease. Consequently, a significant number of renal transplant recipients require a second or subsequent graft. The ability to improve the graft success rate and thereby reduce the number of patients requiring multiple grafts would both improve patient outcomes and increase availability of donor kidneys for primary recipients.

Although advances in renal transplantation such as HLA matching and improved immunosuppressive medication have reduced transplant failure rates further optimization of renal transplant outcomes is necessary to improve both the survival time of the graft and the quality of life of recipients. Techniques derived from the study of artificial intelligence, e.g. artificial neural networks (ANN), offer the ability to better predict graft outcomes after training on a combination of donor and recipient data and thereby to optimize donor–recipient selection. They may also be useful for identification of patients at increased risk of acute rejection and target them for more aggressive immunosuppression.

Examples of ANN applications in organ transplantation include prediction of liver transplant rejection (Hughes et al 2001), prediction of tacrolimus blood levels in liver transplantation (Chen et al 1999), diagnosis of early acute renal allograft rejection and evaluation of complications of renal transplants (Furness et al 1999), predicting cytomegalovirus disease after renal transplantation (Sheppard et al 1999), prediction of pancreas transplant outcome (Dorsey et al 1997) and MHC haplotype matching (Bellgard et al 1998). ANNs can be used to predict delayed renal allograft function and to identify the most important variables in prediction of chronic renal allograft rejection progression (Brier & Aronoff 1996).

To further demonstrate the applicability of an immunoinformatic approach to organ transplantation we performed a study to see whether an ANN could be trained to predict 6 month graft survival (Petrovsky et al 2002). For this we used renal transplant data from the Australian and New Zealand Dialysis and Transplant Registry (ANZDATA). ANZDATA contains clinical donor, recipient and outcome data on all transplants performed in Australia over the last 30 years. After training, a three-layer feed-forward ANN was able to correctly predict 84.95% of successful transplants and 71.7% of unsuccessful transplants thereby demonstrating that an appropriately trained ANN is capable of predicting both successful and unsuccessful renal transplants. The ANN is better in predicting successful than unsuccessful transplants suggesting that the factors that determine graft success may be inherently more predictable than the factors which determine graft failure. We then used an ANN architecture to see if we could predict the type of graft rejection. The ANN-based system correctly predicted 59% of rejection outcomes with respect to the type of rejection. These results indicate that an immunoinformatic approach is extremely useful for predicting renal transplant rejection and could, therefore, be developed into a useful clinical tool to improve transplantation outcomes. The biggest problem in improving renal transplant allocation may, ironically, not be the development of ANN-based prediction systems but rather gaining the acceptance by clinicians of computer-based predictions. However, the accuracy and impartiality of an immunoinformatic allocation system should ultimately be its greatest strength, as this would prevent bias creeping into organ allocation. The technological advances offered by such methods of graft allocation may ultimately benefit the many patients currently awaiting organ transplants.

Human immune disease-gene identification

The majority of common diseases such as cancer, allergy, diabetes or heart disease are characterized by complex genetic traits where genetic and environmental components contribute to disease susceptibility (Hirschhorn et al 2002). Our knowledge of genetic factors contributing to the risk of common diseases is, however, limited. A major goal in the post-genomic era is to identify and characterize disease susceptibility genes and to define strategies to use this knowledge for disease treatment and prevention. The mouse is the most important model organism for the study of human disease genetics, and discovery and validation of potential therapies. Genetic manipulations that can be performed in the mouse include point mutations, gene disruptions, insertions, deletions, or chromosomal rearrangements or random genome-wide mutagenesis (Muller 1999, Zambrowicz & Friedrich 1998). The FANTOM2 project has focused on the functional annotation of 60770 cDNA RIKEN clones by large-

scale, computerized annotation followed by manual curation. Being the most complete picture of the mouse transcriptome to date, the FANTOM2 dataset provides an ideal opportunity for identification of novel mouse orthologues of human genes involved in normal immune function and/or immune disease. We analysed the RIKEN dataset to identify potential novel genes in mouse that are highly similar (70–85% in more than 70% length) to human counterparts that have been described in relation to immune disease and found 14 mouse clones related to rheumatoid arthritis, systemic lupus erythematosus, Crohn's disease and Sjogren's syndrome with nine of the genes encoding autoantigens.

Systemic lupus erythematosus (SLE) is an autoimmune disease associated with impaired humoral and cellular immune responses characterized by a chronic inflammation of the connective tissue, affecting different systems such as joints, kidneys, serous surfaces, and vessel walls (Oelke & Richardson 2002). Antibodies to multiple different self proteins are found in the serum of these patients (Lim et al 2002). Antibodies against DNA, nucleoproteins, histones, nuclear ribonucleoprotein and other nuclear constituents (anti-nuclear antibodies) are found in more than 98% of patient with SLE. Identification of these autoantibodies is of high diagnostic value for SLE. An immunoinformatic search of the RIKEN dataset found novel mouse transcripts similar to the SLE autoantigens, DEK protein, AHNAK, replication protein A and U1 small nuclear ribonucleoprotein C. Rheumatoid arthritis (RA) is a chronic disease characterized by the presence of an inflammatory infiltrate in the synovial capsule inducing progressive destruction of bone and cartilage in the joints. We identified novel mouse proteins similar to L1 retrotransposable elements, small nuclear ribonucleoprotein-associated protein and cAMP-responsive element binding protein, which have all been previously described as autoantigens in RA. We found a mouse transcript similar to Golgin-97, a Golgi complex antigen that is an autoantigen in patients with Sjogren's syndrome (Griffith et al 1997). We also identified a mouse sequence similar to the human zinc finger protein, Cezanne (cellular zinc finger anti-NFκB), a NFκB negative regulator (Evans et al 2001). NFκB is related to multiple inflammatory diseases including rheumatoid arthritis (Muller-Ladner et al 2002). A mouse clone was identified that bore close similarity to NOD2, a member of the Apaf1/Ced4 superfamily of apoptosis regulators that activates the nuclear factor NFκB to enable monocytes in response to bacterial challenges, and is defective in patients with Crohn's disease (Hugot et al 2001). Finally, a mouse clone similar to human HA-1 was identified that may represent a novel minor histocompatability antigen. Minor histocompatibility complexes play a key role in graft versus host (GVH) as well as playing a beneficial role in the graft versus leukaemia response (GVL). Identification of antigenic peptides related to minor histocompatibility molecules responsible for the induction of

GVL and not GVHD is a promising avenue for treatment of allogeneic bone marrow and haematopoietic stem-cell transplantation (Warren et al 1998).

Antigen presentation pathways and their role in human disease

T lymphocytes (T cells) have evolved to be the major effectors of cognate immunity in the vertebrate immune system. T cells possess receptors (TCRs) which, in a highly specific manner, recognize human leukocyte antigen (HLA)-presented peptides on the surface of host cells. HLA molecules bind peptides produced by degradation of proteins. The transporter associated with antigen processing (TAP) is a transmembrane protein responsible for the transport of antigenic peptides into the endoplasmic reticulum where they are then bind to HLA class I. The importance of TAP to the function of the HLA class I antigen presentation pathway is demonstrated by markedly reduced cell-surface HLA class I expression in cells defective in TAP expression (Spies & DeMars 1991). Understanding the pathways of antigen processing and presentation is important for the design of immunotherapeutic drugs and vaccines and for understanding the mechanism behind HLA-associated disease associations.

Analysis of the relationship between TAP binding affinity and HLA class I binding affinity across the full spectrum of HLA alleles is difficult because of the extensive polymorphism of HLA molecules. We addressed the problem by: (a) generating a computational model, (b) combining the initial model with a selected set of laboratory experiments for model refinement, and (c) using the refined model to analyse the functional relationship between TAP and HLA class I molecules (Daniel et al 1998). The working ANN model was used to search for patterns of TAP-binding within sets of HLA-binding peptides. The proportion of HLA-binding peptides with affinity to TAP varied for each HLA class I allele with a range of 15% for HLA-B*5401 to 100% for HLA-B*2703 (Brusic et al 1999). On the basis of these results we hypothesize that HLA alleles constitute two separate classes: those that are TAP-efficient for peptide loading (HLA-B27, -A3 and -A24) and those that are TAP-inefficient (HLA-A2, -B7 and -B8). The strong similarity between the sets of peptides bound by TAP and HLA-B27 suggests close functional co-evolution. Advantages could include increased resistance to infection, cancers or autoimmunity at the individual or species level, but at the price of increased predisposition to ankylosing spondylitis. TAP-inefficient HLA alleles utilize TAP-independent peptide transport pathways to a greater degree. Evolutionary pressures may have selected TAP-inefficient HLA alleles to counter mechanisms evolved by pathogens to evade immune surveillance by blocking TAP-dependent peptide transport. The availability of computer-based models of TAP and HLA interaction is helpful in accelerating research into evolutionary

relationships within the immune system as well as in designing more efficient human vaccines.

Prediction of allergenicity

Allergies constitute the most common cause of chronic illness in industrialized countries, affecting approximately one third of the general population. Clearly there is much to be done to better understand and alleviate this debilitating problem. Given the complexity of the field, immunoinformatics offers great potential to deliver new approaches to allergy management and treatment.

The assessment of protein allergenic potential focuses on three main aspects: immunogenicity, cross-reactivity and clinical symptoms. Immunogenicity refers to the likelihood of an IgE antibody or T cell response to a particular allergen. Studies of B cell and T cell epitopes focus on defining recognition sites on allergens. Cross-reactivity refers to the ability of an IgE clonotype or a T cell clone, which was previously induced by one allergen, to react with another allergen. Studies of stinging insect venom allergens have shown that cross-reactivity between allergens that have less than 70% sequence identity is uncommon (King & Spangfort 2000). Studies of immunogenicity and cross-reactivity have applications in the development of immunotherapies and vaccines.

The number of characterized protein allergens is increasing rapidly. The Allergen Nomenclature Sub-Committee of the International Union of Immunological Societies maintains a list of 'agreed' protein allergens with 360 protein allergens having been classified, as of September 2001 (*www.allergen.org/List.htm*). While this list contains many characterized allergens, it is not exhaustive as unlisted allergens can be found in the literature. Prediction tools focus on functional and structural analysis of genes and proteins and identification of those that have allergenic potential. Such tools have been used for prediction of allergenicity, allergen cross-reactivity and T cell epitopes. Structural bioinformatics helps identify structural properties of proteins, such as secondary structure, or tertiary structure that affect allergenicity largely through affecting the IgE binding sites. However, the common characteristics of allergens such as structural, functional, or biochemical properties that explain their ability to elicit allergic responses are still unclear. Basic sequence analysis methods include the analysis of DNA and protein sequences using sequence comparison, sequence alignment, database searching, or identification of various properties of protein sequences. Homologous proteins usually share 3D structure, implying similar function. The first step in the study of a novel protein is usually a search of public databases for homologous sequences using standard or specialized pairwise sequence alignment algorithms followed by multiple sequence alignment.

Some examples of use of sequence comparison and analysis in the study of allergens include molecular characterization of an allergen group from dust mites (Mills et al 1999), characterization of a superfamily of proteins containing the allergenic lipid transfer proteins (Iyer et al 2001), structural characterization of a rice allergen (Izumi et al 1999), analysis of mutations of pollen allergens (Midoro-Horiuti et al 2001), and identification of characteristic motifs in cat (Ichikawa et al 2001) or cockroach allergens (Yang et al 2000). Secondary structure prediction of a selection of pollen, fruit and vegetable allergens using PredictProtein software helped the identification of very similar structural elements and, in particular, the 'P-loop' region as a common domain of pollen and related food allergens (Scheurer et al 1999). This group of sequences displayed strong allergenic cross-reactivity and the presence of common and specific epitopes. Third generation software, PredictProtein, Jpred2, and PSIPRED accurately predict more than 75% of amino acids that form α helices, β sheets and coils. Structural motif analysis can be useful in studies of specific protein properties, e.g. it successfully revealed the presence of coiled-coil helices in the group 5 allergen Der p 5 from the house dust mite (Liaw et al 2001). Protein allergenicity is determined by 3D structure and structural knowledge can, therefore, help provide insight into the molecular basis of allergenicity. For example, a common structural motif comprising a groove located inside an $\alpha\beta$ motif has been identified within diverse allergens (dust mite allergen Der p 1, cysteine protease papain, lipocalin Mus m 1, and ragweed allergen Amb a 5) (Furmonaviciene & Shakib 2001).

A number of specialized clinical allergy tools have also been developed. One example is the allergen avoidance database tool which, when queried, produces an extensive list of skin care products that do not contain known allergens specific for a given patient (Yiannias & el-Azhary 2000). Another program predicts sensitization to flour allergens in bakers using a stepwise logic regression method (Popp et al 1994). We anticipate that more sophisticated bioinformatics tools will appear in the future to support research, clinical practice and the screening of novel synthetic or genetically manipulated (GM) foods and products. Novel GM foods, nutraceuticals, cosmetics and other products need to be carefully assessed for allergenic safety before they reach consumers. Immunoinformatics currently provides the only cost-effective and efficient means of performing such safety screening, hence its importance to the future of market acceptance of GM products.

Discussion

The preceding examples indicate the power of immunoinformatic approaches to accelerate knowledge acquisition in clinical immunology. Bioinformatics has broad applicability to immunology with uses ranging from defining disease

genes and gene pathways to development of clinical prediction and information systems. Given the potential for bioinformatics to transform immunology research, immunoinformatics provides the key to develop radically new immunology treatments and clinical practices. To achieve this objective will require the development of appropriate immunoinformatics frameworks and standards. In addition, attention will need to be focused on how to ensure that clinical immunologists gain confidence in the computational modelling systems used. Immunoinformatics methods will need to be validated according to minimum clinical as well as experimental standards. An exciting prospect is the development of *in-silico* models of entire systems. There is now the technology available to build a virtual immune system, the construction of which will be one of the major challenges for the next decade.

Acknowledgements

DS is a recipient of a scholarship from the Canberra Hospital Salaried Specialists Private Practice Fund. We would like to thank the RIKEN Institute and the FANTOM2 consortium for enabling us to access the RIKEN data.

References

Bellgard MI, Tay GK, Hiew HL et al 1998 MHC haplotype analysis by artificial neural networks. Hum Immunol 59:56–62

Brier ME, Aronoff GR 1996 Application of artificial neural networks to clinical pharmacology. Int J Clin Pharmacol Ther 34:510–514

Brusic V, van Endert P, Zeleznikow J, Daniel S, Hammer J, Petrovsky N 1999 A neural network model approach to the study of human TAP transporter. In Silico Biol 1:109–121

Chen HY, Chen TC, Min DI, Fischer GW, Wu YM 1999 Prediction of tacrolimus blood levels by using the neural network with genetic algorithm in liver transplantation patients. Ther Drug Monit 21:50–56

Daniel S, Brusic V, Caillat-Zucman S et al 1998 Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. J Immunol 161:617–624

Dorsey SG, Waltz CF, Brosch L, Connerney I, Schweitzer EJ, Bartlett ST 1997 A neural network model for predicting pancreas transplant graft outcome. Diabetes Care 20:1128–1133

Evans PC, Taylor ER, Coadwell J, Heyninck K, Beyaert R, Kilshaw PJ 2001 Isolation and characterization of two novel A20-like proteins. Biochem J 357:617–623

Furmonaviciene R, Shakib F 2001 The molecular basis of allergenicity: comparative analysis of the three dimensional structures of diverse allergens reveals a common structural motif. Mol Pathol 54:155–159

Furness PN, Kazi J, Levesley J, Taub N, Nicholson M 1999 A neural network approach to the diagnosis of early acute allograft rejection. Transplant Proc 31:3151

Griffith KJ, Chan EK, Lung CC et al 1997 Molecular cloning of a novel 97-kd Golgi complex autoantigen associated with Sjogren's syndrome. Arthritis Rheum 40:1693–1702

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K 2002 A comprehensive review of genetic association studies. Genet Med 4:45–61

Hughes VF, Melvin DG, Niranjan M, Alexander GA, Trull AK 2001 Clinical validation of an artificial neural network trained to identify acute allograft rejection in liver transplant recipients. Liver Transpl 7:496–503

Hugot JP, Chamaillard M, Zouali H et al 2001 Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature 411:599–603

Ichikawa K, Vailes LD, Pomes A, Chapman MD 2001 Identification of a novel cat allergen — cystatin. Int Arch Allergy Immunol 124:55–56

Iyer LM, Koonin EV, Aravind L 2001 Adaptations of the helix-grip fold for ligand binding and catalysis in the START domain superfamily. Proteins 43:134–144

Izumi H, Sugiyama M, Matsuda T, Nakamura R 1999 Structural characterization of the 16-kDa allergen, RA17, in rice seeds. Prediction of the secondary structure and identification of intramolecular disulfide bridges. Biosci Biotechnol Biochem 63:2059–2063

King TP, Spangfort MD 2000 Structure and biology of stinging insect venom allergens. Int Arch Allergy Immunol 123:99–106

Liaw SH, Chen HZ, Liu GG, Chua KY 2001 Acid-induced polymerization of the group 5 mite allergen from Dermatophagoides pteronyssinus. Biochem Biophys Res Commun 285:308–312

Lim Y, Lee DY, Lee S et al 2002 Identification of autoantibodies associated with systemic lupus erythematosus. Biochem Biophys Res Commun 295:119–124

Midoro-Horiuti T, Goldblum RM, Brooks EG 2001 Identification of mutations in the genes for the pollen allergens of eastern red cedar (Juniperus virginiana). Clin Exp Allergy 31:771–778

Mills KL, Hart BJ, Lynch NR, Thomas WR, Smith W 1999 Molecular characterization of the group 4 house dust mite allergen from Dermatophagoides pteronyssinus and its amylase homologue from Euroglyphus maynei. Int Arch Allergy Immunol 120:100–107

Muller U 1999 Ten years of gene targeting: targeted mouse mutants, from vector design to phenotype analysis. Mech Dev 82:3–21

Muller-Ladner U, Gay RE, Gay S 2002 Role of nuclear factor kappaB in synovial inflammation. Curr Rheumatol Rep 4:201–207

Oelke K, Richardson B 2002 Pathogenesis of lupus. Arthritis Rheum 47:343–345

Petrovsky N, Brusic V 2002 Computational immunology: the coming of age. Immunol Cell Biol 80:248–254

Petrovsky N, Tam S, Brusic V, Russ G, Socha L, Bajic V 2002 Use of artificial neural networks in improving renal transplantation outcomes. Graft 4:6–13

Popp W, Wagner C, Kiss D, Zwick H, Sertl K 1994 Prediction of sensitization to flour allergens. Allergy 49:376–379

Scheurer S, Son DY, Boehm M et al 1999 Cross-reactivity and epitope analysis of Pru a 1, the major cherry allergen. Mol Immunol 36:155–167

Sheppard D, McPhee D, Darke C et al 1999 Predicting cytomegalovirus disease after renal transplantation: an artificial neural network approach. Int J Med Inf 54:55–76

Spies T, DeMars R 1991 Restored expression of major histocompatibility class I molecules by gene transfer of a putative peptide transporter. Nature 351:323–324

Warren EH, Gavin M, Greenberg PD, Riddell SR 1998 Minor histocompatibility antigens as targets for T-cell therapy after bone marrow transplantation. Curr Opin Hematol 5:429–433

Yang CY, Wu JD, Wu CH 2000 Sequence analysis of the first complete cDNA clone encoding an American cockroach Per a 1 allergen. Biochim Biophys Acta 1517:153–158

Yiannias JA, el-Azhary RA 2000 Contact Allergen Avoidance Program: a topical skin care product database. Am J Contact Dermat 11:243–247

Zambrowicz BP, Friedrich GA 1998 Comprehensive mammalian genetics: history and future prospects of gene trapping in the mouse. Int J Dev Biol 42:1025–1036

DISCUSSION

DeLisi: The Human Genome Project is not a good model for what we want to achieve. A better model is Al Gilman's project mapping signalling pathways, the Alliance for Cell Signaling. The human genome project was successful because it was presented as an engineering project: the goals were to develop the high-throughput technologies required for sequencing the human genome, and then to sequence it. The Virtual Immune System is a scientific project. The issue is how we translate scientific interest into government policy that is going to lead to US$10–15 million a year in funding for such a project. NIH still has more money than it knows how to spend intelligently, and one can think about a US$10–15 million effort if it is well defined and articulated, with some good end points.

Perelson: We have been thinking about whether or not this is feasible for many decades. One of the problems in doing modelling is that we want it to be related to the data. What is interesting about immunology is that different experimental groups have focused on different aspects of the immune response, and with different model organisms. From what we can tell, there is no single response to a relevant pathogen that has been completely characterized. People tend work on model antigens such as myoglobin or hen egg lysozome, which are easy to obtain and which give rise to large, easy to measure responses. There has been little integration in the immunological community to take a real antigen, such as a pathogen, and then completely characterize T cell, B cell and cytokine responses. A number of years ago we were considering getting together a group similar to Max Delbruck's phage group, with interests in different aspects of immunology, and have them work on one model antigen to start building an integrated picture of how the different arms of the immune system work together. Our hope was that this effort would collect data that would drive a model integrating different aspects of the immune response. Initially, people were intrigued and we had about five different labs that were going to try. It eventually fell apart because everyone had expertise working with their particular antigen. To characterize another antigen people were willing to expend 5% effort, but they weren't willing to make a major change in their career or laboratory's expertise to move to someone else's antigen or system. Perhaps now the time is right to revisit such an endeavour. There is a much larger community of people interested in quantitative information. There is a large amount of money coming into studying immune responses related to bioterrorism threats. With funding and incentive it may become possible to build a large-scale model of an immune response to a pathogen, but the current state of modelling is still rather primitive in many respects and a large amount of work will need to be done. The early models of immune responses were predator–prey models. They would say that there is

some antigen that is either growing or not growing, and the antigen level would drive immune responses in a concentration-dependent manner. We have clever physicochemical models, at the level of B cells, of how receptor cross-linking might occur and how this might generate signal transduction. Similarly, there are now new models of the immunological synapse and the early events in T cell activation (cf. Coombs et al 2002). But one of the things that we have learned in the last few years is that the immune system might not work by simply monitoring the antigen concentration and having its response driven by that. We have learned from *in vitro* models that much of the response may be pre-programmed. There are proliferative responses that take place after a system has been pulsed with antigen, and if the antigen is removed the response still continues for a number of days (cf. van Stipdonk et al 2001, Badovinac et al 2002).

We are still in a primitive state with regard to fully understanding the signals that regulate T cell proliferation and differentiation. The same is true of B cells. So I think there has to be a combined effort where experimental and modelling groups work together and carefully elucidate the kinetic details of responses in model systems. We also need to look at the responses to multiple antigens that occur in real infections, and consider cross-reactivity and competing cytokine signals, for instance. To build an accurate large scale model of the immune system is an effort that will take many years, but unless we start it will never be achieved.

One further area that relates to immunoinformatics is that people who are trying to characterize responses in T cells have a technology called the immunoscope that measures the length of the CDR3 region of the β chain of the T cell receptor (Pannetier et al 1995) rather than the full genetic sequence. Labs collecting huge amounts of data of this sort are trying to characterize the diversity of the Vβ families of T cell receptors, and how they change during immune responses. I don't know what is happening to all these data. It would be useful to put this information in databases so these data can be universally accessed.

Brusic: Nikolai Petrovsky's optimism stemmed from successful linking of predictions with subsequent experimental validations of TAP binding peptides. We found someone who was willing to do a TAP binding study with 100 randomly generated peptides. By using such an approach we defined a strategy for defining highly accurate models. This is the main problem: it is very difficult to get experimental people to invest in randomly generated peptides where they don't clearly see what is coming out.

Perelson: There are two issues. One is building models of particular systems. If one wants to build models of TAP and MHC presentation, there are whole groups of people who would be willing to do this. What Nikolai is proposing is something much broader: trying to build a model of the whole immune response. This involves getting people signed on who are not only interested in antigen

presentation, but also the functional T cell response, the antibody response and possibly people interested in vaccines. Another aspect of global modelling is that depending on the organism you focus on, immune responses take different characters. This is because the response is occurring within a host, and the host can choose to mount a cell-mediated response, a humoral response or both. The organism also may be trying to subvert the immune system such as by interfering with antigen presentation. Thus the data one collects and the models that are developed may depend on the particular host–antigen system being studied.

Petrovsky: What I was proposing is somewhat different to what people have proposed before in that I am advocating a modular approach. If you can develop a model of TAP binding that is close to 99% accurate, which we believe we have done, you can then combine this with well validated models of MHC binding which are also highly accurate. By combining these two modular components we now have a significant part of the MHC Class I antigen presentation pathway sorted out. You can then proceed to model other components such as the proteosome or T or B cell recognition. In essence, we want to progressively build a complete model of the immune system so we can put ideas in and see what comes out the other end. The responses of the model can then be compared to observations of responses in whole organisms.

Rammensee: I would like to comment on the two nice models you proposed. One, the mutation model, was a black box model and was very successful. The other was model-based predictions building on blocks. The problem with this building block-based approach is that you are making the assumption that these two blocks do everything and there is no additional influence. Usually, though, nature is more complex. This causes a problem for the building block-based model. Your example just missed the state between TAP and MHC, which is now well established. This is where the trimming enzyme which shortens the peptides comes in. The peptides are made shorter by aminopeptidase activity to nine amino acids, so you don't need the peptides coming in at exactly the right length. Since this is missing in your prediction mark, it calls into question the whole thing. You need to include this new building block of the specificity and activity of the trimming peptidase and perhaps then the whole thing will be more accurate.

DeLisi: There are models that do exactly what you are saying.

Rammensee: You have to know these steps from the experiments. Otherwise you are in the black-box business.

Lybrand: I don't think this is necessarily the case. Perhaps it is a difference in perspective of how you do modelling and exactly what modelling is. You can take the kind of approach that I think Nik is outlining and try to couple these different levels of models together. In many cases, what the overall mathematical model will help reveal to you is where you have missed some steps. Initially,

models define limiting behaviours and you have to begin to refine them to come to a better accord of what you can measure. It can be used to steer you back to areas where you need more detailed experimental investigations.

Rammensee: In his TAP B27 example Nikolai Petrovsky did not conclude there was something missing.

Petrovsky: I agree entirely with the comments made by Terry Lybrand, namely that the model defines limiting behaviours and helps us to identify missing components. Because we couldn't conceive of a whole class of MHC class I alleles that have no capacity to bind peptide because it can't get in to the Class I pathway via normal TAP transport we started to look for the missing links: in other words, how do these other peptides get into the Class I pathway if it is not through TAP transport? Either, as Hans Georg says, it is because larger peptides are being transported by TAP and then they are being trimmed once they are in the ER, or perhaps these peptides without a TAP binding motif are from viruses which are expressed in the ER and would be anticipated to be loaded into MHC class I molecules like HLA-A2 that bind peptides that don't have a TAP binding motif. When we did a literature search, we found that there does seem to be a bias whereby the HLA molecules that don't have similar binding to TAP seem to be important for presenting viral and signal peptides. Models aren't meant to be perfect, but they do alert you to incongruous results that make you then look experimentally to find the missing links.

Rammensee: This missing link — the trimming peptidase in the ER — has recently been published (Serwold et al 2002).

Petrovsky: This could easily be built into the model and I agree this should be the next step.

Bernaschi: I think we should distinguish between modelling that attempts to explain the result of experiments and classical modelling, which tries to find the smallest possible set of differential equations which describes the result and can be useful for understanding the experiment. It could take many years for us to find the right set of equations for the immune system. A completely different approach to modelling is to try to simulate what we already know. Let me try to explain this point by means of an analogy. We don't know everything about the laws of physics in fluid dynamics, and we don't know much about turbulence. But we have very detailed flight simulators, which are very useful for engineering. We should adopt a slightly different approach to immunoinformatics. On the one hand we should try to amass as much data as possible to provide a classic scientific model. On the other hand we should start as soon as possible to use the information we have to simulate the immune system. They are quite different stories. Of course, any simulator must be validated. This is always done in fluid dynamics. There is a lot of working code, but no one would seriously suggest flying an aeroplane that was simply designed by computer. There are wind tunnels that are used for

experiments, but the use of computer simulations saves a lot of money. The same story could be true for immunology. Another point raised by Vladimir Brusic concerned the necessity for joining different scales. We are already able to reproduce many effects at the cellular level. We have less information for the intracellular level. These are different challenges. We now have enough information, however, to write something that is useful and will be a starting point for the real immune system simulator. I don't know whether we will ever be able to assemble a complete picture of a system as complex as the immune system, but I am pretty sure it is useful to start with the information we already have.

DeLisi: A simulator would drive the experiments. It would have parameters that we would know from experiments. But experimental parameters at one level of modelling are predictable at a deeper (more fundamental) level. You want to do experiments that give you the feeling of the parameters. For example, take cell adhesion. You can measure the thermodynamic and kinetic binding constants between particular cells under particular conditions, and they would allow you to make predictions for those cells under those conditions, e g. how their interaction rate changes with concentration, temperature receptor density, etc. That would be helpful, but not very general. A theory of the binding constants would allow you to extend the range of predictions enormously. For example, we might want to predict adhesive properties knowing only the alleles of the relevant genes. From those we would determine the protein sequences and from sequence determine structure and from structure determine free energies in terms of surface amino acids (the docking problem). What is a parameter at one level becomes detailed experiments at another. I agree with that idea; in theory a simulator would drive experiments at different levels. We are clearly never going to understand everything, so the goal here is to be very concrete about what we really want to accomplish with our resources. If we had US$15 million a year, what would we do with this money towards that goal? There is a venture called the Virtual Human Project. Its goal is to develop organ-level models and then to begin to connect them. The immune system is one such organ. Other than the CNS, it is probably the most complex. Modelling it certainly fits with a lot of the culture that is developing at least in the USA.

Gulukota: We have talked a great deal about data and a need for a data repository. One concrete idea could be that one institute hosts a PDB-type database that is well annotated with respect to the techniques that are used. With regard to neural networks, and people's reluctance to use them because it is a black box, if you have a black box that you believe works almost 100% of the time, you can do pretty much infinite calculation with it. This creates an infinite database because you can place all your calculations into a repository. Also, the neural net doesn't have to be a black box in terms of explaining for two reasons. First, there are

techniques for differentiating neural nets which allow you to identify which input components contribute most to the analysis. Second, with a large enough data repository such as one created above, you can apply any other statistical tool to recognize more intuitive patterns. Finally, I have a question for you. You mentioned that you were looking at the success and failure of transplantation. I guess you clarified this a little by saying you were looking at success 6 months out. Is that really binary?

Petrovsky: 'Failure' at any time point is an absolute end point in that it means that the graft has failed and the patient is back on dialysis. 'Success' is more semantic in that when we talk about 'success' at six months, we just mean that the graft is still functioning at that time point. We plan to extend the study to see whether we can similarly predict transplantation outcomes at one year, two years etc.

Gulukota: From a scientific point of view, is it useful to look at some other description of success or failure, such as the histology of the graft? Clinical outcome is important, but is it important to correlate the black box with something a little more concrete?

Petrovsky: Clinicians like studies with solid end points. In transplantation the hardest endpoint is a graft that is functioning or one that has failed. A clinical endpoint like this is more valid than measuring something such as serum creatinine or graft histology, which may be good or bad but may not actually reflect the final graft outcome.

Beck: How much are these data skewed by medication? After the initial success, some people might be on strong medication and some might be on no medication at all?

Petrovsky: What we have done is make the predictions based solely on pre-transplant data and we don't incorporate any post-transplant data. The goal is to produce a method for allocating organs for transplantation more effectively. The amazing thing is that we don't know what medication a patient will receive post-transplant and yet we are still able to predict the graft outcome at a reasonable level of confidence. This would suggest that at least at 6 months post-transplant medication alone cannot fully compensate for other intrinsic factors influencing graft outcome, for example the HLA types of the donor and recipient.

Beck: The post-transplant data would be useful. You rarely get the perfect match. If you get a second-best match, it may be that this will work better with a particular combination of drugs but not another.

Brusic: It is difficult to get clean transplant data because of the known problem of patient non-compliance in taking drugs.

Petrovsky: We certainly plan to address these sorts of questions with the system. You could even predict how much drug a patient is likely to require in order to have a successful graft. To go into finer detail and answer such questions, more data are required to take into account the complexity. For example, one of the things

that does influence outcome strongly is the surgeon performing the transplant. The neural network understands this. If we take these data out, the performance of the model falls. The beauty is that the network is a black box: one of the things that people were very worried about is that the network would start to point fingers at particular surgeons. The neural network in fact does this in a subtle way, by weighting the outcome downwards if it knows a particular surgeon with a history of poor outcomes is proposed to do a particular transplant, but at the same time it is taking into account many other variables so it is never possible to say that the outcome is predicted to be bad just because of a particular surgeon.

Rammensee: Would you assume that your model would get better the more parameters you put in the black box, without knowing what the influence would be?

Petrovsky: Exactly. The neural network will discard irrelevant data, or those data that don't contribute to the prediction, hence you can put in as many data as you like and let the network sort out what is useful. This is different to many other systems where irrelevant or fuzzy data may detract from the prediction performance.

Rammensee: That's not very scientific, but it seems to work.

Gulukota: I wasn't saying that the end point is not scientific. I am suggesting that if we want to understand the rejection process and see what will happen further down the line, it might be more useful to look at some other markers.

Roth: That's a fair point. We had clinical studies where pathologists assessed kidney biopsies for chronic rejection. If you ask two pathologists they give you two different opinions about the state of the graft. Not even the pathologists have a clear opinion of how to assess rejection.

Kesmir: I disagree with the idea of neural networks being a black box. Just because the mathematics behind it contains some partial differential equations doesn't make something a black box. All the parameters, i.e. weights, of the network can be analysed. Then you can understand why it makes certain predictions.

Rammensee: The collection of data is the black box, as I understand it.

Kesmir: It is also possible to analyse the data. The worst black box is a mouse! When we measure T cell responses, we don't know what they are due to.

Flower: You won't find many immunologists or clinicians who are prepared to sit down with a neural network, take it to pieces and try to understand how it works. It would be naïve to think otherwise. One must recall that the goals and interests of distinct groups of scientists do differ. You will not even find many computational chemists who are prepared to take a neural network to pieces. It is perhaps only computer scientists who might be prepared to disassemble a neural network.

Kellam: This may change in the near future so we would get more cross-disciplinary research collaborations. What I don't understand is the statement that it is not scientific to use a 'black box' analysis method if you assimilate a lot of data, from a hypothesis and test it. How is this any different from taking

the data, running them through a defined prediction method and finding that this fits the data?

Gulukota: I was merely suggesting that looking at a different end-point might give better predictive value over a longer time scale. You might even be able to predict how long the graft will survive.

Petrovsky: We can certainly predict how long the kidney graft will survive if we train on the appropriate data. If we incorporate post-transplant data the system becomes more powerful at predicting outcome but would no longer be useful for allocating cadaver kidneys as this can only be done based on data that is available pre-transplant. The trouble is, once the transplant has already occurred you can no longer go back and change things.

Roth: Post-graft data can help to tailor the immunosuppressive regime.

Petrovsky: Absolutely. Obviously a neural network approach can be used to assist many separate clinical decisions in transplantation.

Gulukota: But you could look at post-transplant data for training purposes and then use these data to train the neural network. Then for the testing set you only have data coming from before the transplant is done.

De Groot: I have a question for Nikolai Petrovsky and Vladimir Brusic, because they seem to have two pieces of information that I don't have: namely, that there is a model of TAP that predicts an inverse relationship between what ends up as an epitope and what binds to TAP. Hans-Georg Rammensee also mentioned that there was a new finding published in *Nature*. How would you explain what you observed? To me as an immunologist it is very important to understand that relationship. Do you feel that there is a second path by which the peptides are getting in to the ER?

Brusic: There are at least four or five different pathways, but TAP is the major one. It is responsible for approximately 90% of antigen transport.

De Groot: You are presumably using B7, A2 and A3 to model, and the model predicted that these would bind to TAP. Yet they do not, and the ones that are poor binders by your prediction method are not the ones that are epitopes.

Petrovsky: I agree that the predictions made by the model are very interesting even if at first sight they appear somewhat paradoxical. What the model predicted was that less than 1% of peptides that bind HLA-A2 would bind to TAP. But the truth is that TAP transports millions of peptides, so even if the proportion that bind to HLA-A2 is less than 1%, if you multiply this by the millions of peptides available, there will still be many peptides transported by TAP that will bind A2, but at a much lower frequency than for example is the case for HLA-B27 which is predicted to bind with high affinity close to 100% of peptides transported by TAP.

Rammensee: The assumption from the model is wrong. It assumes that 9-amino-acid peptides are transported but not 10-amino-acid peptides, and this has now been shown to be wrong.

Margalit: It is possible that the HLA-A2 peptides are cleaved with an N-terminal extension and they are transported by TAP, like the 14-mer.

Brusic: Another explanation for this was that longer peptides are transported in the context of HLA-A2, and it takes time for post-processing in endoplasmic reticulum. If B27 peptide is processed in the cytosol to the optimum length then it can be transported and presented very efficiently.

Petrovsky: You might anticipate that optimal length peptides will be more efficiently processed but I agree that longer peptides may contain HLA-A2 binding motifs. However, by definition all TAP transported peptides will also have a B27 binding motif so there will still be competition between HLA molecules for which gets to capture the TAP transported peptides and the model would predict that B27 if present is usually going to win the race to capture most TAP transported peptides.

De Groot: If you were to try to create something in the cytosol that is an A3 epitope, should you make it a 14-mer or a 9-mer?

Petrovsky: Ideally, you should make it a 14-mer with a TAP binding epitope inserted into a A2 epitope, or linked to it.

De Groot: This is an important issue for those of us who want to apply this. The other possible interpretation might be that it is more important to predict what binds to the HLA than to worry about TAP.

Petrovsky: I don't think we should simply ignore TAP binding as it is clearly an important route for antigens to get into the class I pathway. With our current knowledge we can now model peptides to maximize their chances of both being transported by TAP and binding to any particular HLA molecule. This is a major advance on just being able to model HLA binding.

Margalit: I recently read a review by Alfred Goldberg that suggests the contrary (Goldberg et al 2002). He said that it is better for peptides to have an N-terminal extension because they are being degraded fast by aminopeptidases present in the cytosol. If they are longer the extension will be chopped and then they will have a chance to be transported to the MHC at the right size.

Brusic: How hard are these data?

Rammensee: There are now data that even 9-mers transported inside the ER have been attacked by peptidases before they can bind (Serwold et al 2002).

Brusic: The arginine is a primary anchor at position 2 for peptides that bind to HLA-B27. It is also an ideal place for cleavage.

Rammensee: Aminopeptidases usually don't distinguish between residues unless there is a proline there, which they don't like.

Borras-Cuesta: You seem to predict most of the peptides that are transported by TAP, which is very interesting. Suppose you have an antigen that is 300 amino acids long. Then you screen the possible 9-mers or 14-mers. How many peptides end up being predicted as transported?

Petrovsky: Less than 3%. If you look at random peptides, the number predicted to have high TAP binding affinity is down around 1%.

Borras-Cuesta: How does that correlate with the peptides that are presented in any given model? If it is that restrictive, it would be marvellous.

Petrovsky: We generated completely random peptides and the predictions of the neural network model were shown to be 99% accurate, so we are not missing many peptides at all. Hence if you just look at 9-mers the system is highly restrictive, and at least as restrictive as HLA binding. The difficulty though, is knowing how many peptides are transported by TAP as longer fragments and then cleaved. This would potentially reduce the level of restriction. The interesting thing about TAP is that it appears to be the major gate keeper to the class I pathway and therefore any level of restriction is going to be important. Once peptides get into the ER then they are going to have a large choice of different HLA molecules to bind to so the restriction is going to diminish.

Margalit: In your paper you also tested longer peptides, and you saw that only the first three or four amino acids have the largest contribution (Daniel et al 1998). These considerations should be incorporated in more advanced versions of the predictive scheme that will also deal with longer peptides.

Petrovsky: When you start building these models you keep them relatively simple because you are not sure they are going to work in the first instance. Having shown they work really well, the answer is then to introduce more complexity, such as different lengths of peptide. Again, this is a preliminary model. We certainly plan to model longer peptides in due course.

References

Badovinac VP, Porter BB, Harty JT 2002 Programmed contraction of $CD8^+$ T cells after infection. Nat Immunol 3:619–626

Coombs D, Kalergis AM, Nathenson SG, Wofsy C, Goldstein B 2002 Activated TCRs remain marked for internalization after dissociation from pMHC. Nat Immunol 3:926–931 [Erratum in Nat Immunol 3:1109]

Daniel S, Brusic V, Caillat-Zucman S et al 1998 Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. J Immunol 161:617–624

Goldberg AL, Cascio P, Saric T, Rock KL 2002 The importance of the proteasome and subsequent proteolytic steps in the generation of antigenic peptides. Mol Immunol 39:147–164

Pannetier C, Even J, Kourilsky P 1995 T-cell repertoire diversity and clonal expansions in normal and clinical samples. Immunol Today 16:176–181

Serwold T, Gonzalez F, Kim J, Jacob R, Shastri N 2002 ERAAP customizes peptides for MHC class I molecules in the endoplasmic reticulum. Nature 419:480–483

van Stipdonk MJ, Lemmens EE, Schoenberger SP 2001 Naive CTLs require a single brief period of antigenic stimulation for clonal expansion and differentiation. Nat Immunol 2:423–429

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

Immunoinformatics in personalized medicine

Kamalakar Gulukota

gvk bioSciences Private Limited, #210, 'My Home Tycoon', 6-3-1192, Begumpet, Hyderabad 500 016, India

Abstract. Diagnosis of human disease has been undergoing steady improvement over the past few centuries. Many ailments that were once considered a single entity have been classified into finer categories on the basis of response to therapy (e.g. type I and type II diabetes), inheritance (e.g. familial and non-familial polyposis coli), histology (e.g. small cell and adenocarcinoma of lung) and most recently transcriptional profiling (e.g. leukaemia, lymphoma). The next dimension in this finer categorization appears to be the typing of the patient rather than the disease i.e. disease X in person of type Y. The problem of personalized medicine is to devise tests which predict the type of individual, especially where the type is correlated with response to therapy. Immunology has been at the forefront of personalized medicine for quite a while, even though the term is not often used in this connection. Blood grouping and cross-matching (for blood transfusion), and anaphylaxis test (for penicillin) are just two examples. In this paper I will argue that immunological tests have an important place in the future of personalized medicine. I will describe methods we developed for personalizing vaccines based on MHC allele frequencies in human populations and methods for predicting peptide binding to class I MHC molecules. In conclusion, I will argue that immunological tests, and consequently immunoinformatics, will play a big role in making personalized medicine a reality.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 43–56

We are witnessing the slow unfolding of a paradigm shift in the practice of medicine. Hitherto, only a small number of specific attributes of patients like age, gender, pre-existing conditions were considered important in the practice of medicine — in diagnosis and therapy. Many other attributes of a person, such as their genetic profile, were relatively unimportant (except perhaps in forensic medicine). Now, with the advent of pharmacogenomics, we stand poised to exploit a great deal more information about patients to help with diagnosis and also to help make the decision on what therapy to prescribe. The new paradigm is often referred to as personalized medicine (PM). In this paper I will describe a few scenarios of how immunological data analysis can help to personalize medicine.

Definition of personalized medicine

It is important to define PM not in terms of specific technologies (e.g. genetic haplotyping, transcriptional profiling) but in terms of the results we wish to attain. This is because most technologies which could be applied to PM are still being developed and it is far from clear which (if any) will eventually succeed. Thus a good definition of the problem is as follows.

> 'PM aims to develop diagnostic tests which classify a patient (rather than a disease) in such a way that the classification is correlated with response to therapy.'

The most useful tests will be purely predictive, i.e. they can be employed before any therapy is started and their results used to decide on a therapeutic course of action. Then there are semi-predictive tests which can be employed only after drug administration has begun but whose results become available before the effect of the drugs (effective, ineffective or adverse effect) becomes clinically obvious. These semi-predictive tests can also be useful in that they help make quicker decisions about changing the drug regimen where necessary.

Pharmacogenetic approaches

The most prominent way to approach personalization of medicine has been to look for correlation between human genetic variation and drug response. Since the largest number of variations in the human genome is through single nucleotide polymorphisms (SNPs) (International Human Genome Sequencing Consortium, Lander et al 2001), the obvious initial approach is to examine whether there are any SNPs in the coding region of the gene encoding the drug target. Then, the next step is to see if the presence of these SNPs correlates with drug response. However, SNPs in non-coding regions such as untranslated regions, introns or promoter regions might also be valuable because these might affect regulatory elements. Also, only a portion of all human SNPs are known. Therefore it is quite possible that an important SNP within the coding region is not yet known but a known SNP in the non-coding region might be in linkage disequilibrium with it.

Another very important set of SNPs (from a drug response point of view) would likely affect coding or non-coding regions of genes encoding drug metabolizing enzymes (Hall 2002) as these could have a profound impact on pharmacokinetics.

These two sets potentially make a large number of possible SNPs to examine. Nevertheless, it has been possible to use prior knowledge about the function of the enzyme thiopurine S-methyltransferase (Krynetski & Evans 2000) in the metabolism of the drug mercaptopurine to narrow down this list of SNPs to

identify successfully genotypes associated with toxicity to mercaptopurine (Evans et al 2001).

Genetic variation considered above has a direct effect on drug response since the SNPs are located in the genes for the targets or for the metabolizing enzymes of the drug. However, it is possible to further enlarge this already large list by including SNPs that are located in genes which share a common metabolic pathway with the above 'direct' genes. If so, the effect of each of these indirect SNPs might be small but combinations of them might still produce significant drug response phenotypes (Drysdale et al 2000).

Such expansion (to other related genes) pretty quickly brings it close to a genome-wide SNP analysis. The estimates of the number of SNPs needed to do a genome-wide haplotype vary from 180 000 to 600 000 (see Judson et al 2002, for a review). Most of these SNPs represent genetic variation which will probably have no bearing on the particular drug response in question. Thus correlating a haplotype based on a large number of SNPs to drug response phenotypes becomes a highly statistical exercise often requiring very large sample sizes. For these reasons, few clear correlations have yet been established between SNP haplotypes and drug response.

Immunological approaches

It is surprising that immunological approaches are not prominently used in PM because the earliest and most robust known tests which fit the PM definition (though they were not called PM) are from immunology. Examples include penicillin skin test for allergy, blood grouping for transfusion and HLA typing for transplantation. I will present two approaches I have been involved with which aim to make PM a reality: personalized vaccines and computational approaches to predicting MHC-peptide binding.

Personalized peptide vaccines

Vaccines made of short peptides have at least three advantages over the traditional (full protein or whole organism) vaccines. First, they can be chosen in such a way as to be 'broad spectrum', i.e. effective against a large number of strains of the pathogen. Second, their manufacture is easier to scale-up to production quantities than that of long proteins. Third, it can be expected that they will eliminate the rare adverse side effects that afflict traditional vaccines.

Host and pathogenic genetic variation. Considering the well established principles of immunology, peptides constituting vaccines will need to:

(1) be a part of the pathogen's proteome,
(2) bind one of the MHC molecules in the host, and
(3) trigger a T cell response in the host.

While condition (1) appears trivially obvious it could present practical difficulties, especially in the case of rapidly mutating pathogens like HIV. In these cases, the pathogen population infecting a single host is usually very non-homogeneous resulting in immune escape of some sub-population or other of the pathogen. In peptide vaccine design there is an effective solution to this problem: look for conserved regions within the pathogen's proteome and choose peptides exclusively from these regions. Notably, this solution is not feasible when considering whole proteins as vaccines, since it is unlikely that a whole protein will be conserved across many strains of a pathogen. Thus peptide vaccines could be designed to be 'broad spectrum' over a large number of strains of a pathogen.

While condition (1) deals with pathogenic polymorphism, condition (2) raises the problem of polymorphism in the host since the MHC locus is very polymorphic. Because vaccines are typically expected to be effective over a whole (host) population, which MHC allele should a peptide vaccine bind? One approach to solve this is to look for promiscuous peptides, i.e. peptides which bind multiple MHC alleles. However, the universe of possible peptides has already been shrunk to just the conserved regions of the pathogenic proteome by condition (1). To find peptides that bind any MHC allele within these could already be a difficult exercise. To find one that is promiscuous would be doubly difficult. An alternate approach might be more promising: design the vaccine as a cocktail of peptides such that a majority of the host population has an MHC allele which binds at least one of the peptides in the cocktail.

One could also write other conditions (for example, related to proteasomal cleavage of antigens) but I have left these out for the purposes of this presentation. Alternately one could view these conditions as being subsumed in the loosely defined condition (3).

How many alleles for 90% coverage? One approach we took (Gulukota & DeLisi 1996) to address the MHC polymorphism issue was to ask the question: what is the minimum number of HLA alleles needed, in order to 'cover' a prespecified proportion (say 90%) of a given ethnic group? 90% coverage implies that 90% of people in the ethnic group have at least one of the chosen alleles. This appears to be a trivial problem given the estimated frequencies of various HLA alleles; we used the data tabulated by Imanishi et al (1993). However, a complication arises due to linkage disequilibrium between alleles: the frequency of two alleles occurring together (joint probability) is not the product of frequencies of individual alleles.

We devised a solution to tackle the linkage disequilibrium issue and tabulated the alleles required to cover various ethnic groups at 90% (see Table 2 in Gulukota & DeLisi 1996). Three to six alleles were sufficient to cover most ethnic groups at 90%. The relatively homogeneous groups such as the Southern Han Chinese could be comfortably covered with three alleles while the more diverse populations like those from Africa required up to six alleles.

How is this personalization? The chosen coverage (arbitrarily, 90%) is an upper bound on the efficacy of the vaccine. This appears like a major problem because this procedure will handicap the vaccine from the beginning with this theoretical upper limit. But three considerations somewhat mitigate this concern. First, while a vaccine efficacy under 90% appears rather poor, for problematic viruses like HIV, 90% efficacy will be a major advance. Second, 90% can still induce substantial herd immunity and help arrest the spread of infection.

Third, there is this argument, connecting this up with PM: mixing cocktails which have broad (90%) coverage of whole populations appears more like 'ethno-selection' than like personalization. But taken to its logical conclusion this procedure can in fact be a sound basis for personalizing vaccines. For example, mixing up the cocktail of peptides is something that can be achieved as a formulation in the pharmacist's office. Thus, multiple cocktails could be used to cover the whole population. And for persons especially at risk, a prior diagnostic test to determine their HLA haplotype could be used to determine which cocktail would be most effective for them.

Pointedly, such a use of HLA haplotyping clearly fits the PM definition.

MHC–peptide binding prediction

Having chosen what HLA alleles are important for which ethnic groups, we set out to develop computational methods for predicting peptides binding to these alleles. HLA-A2 is a very prevalent allele at the A locus among almost all ethnic groups, although its actual frequency varies; A2 turns out to be a component of all 90% coverage peptide cocktails we examined. We started with A2 (Gulukota et al 1997) as our first test allele.

We developed two complementary methods for predicting peptide binding to HLA-A2: one based on artificial neural nets (ANNs) and the other called polynomial method, is based on the independent binding of sidechains (IBS) assumption (Parker et al 1994).

For both methods the raw data were generated using the IC_{50} method (Kast et al 1994) by our co-authors. We believe this consistency among the data is important when looking for statistical correlations. Mixing data from a variety of methods could lead to patterns that are complicated by the superposition of different biases peculiar to the experimental techniques employed.

Since IC_{50} measurements yield a continuous variable, it is important to decide on a resolution for the prediction i.e. are we looking to predict actual IC_{50} or some coarser representation of it (like 'strong binder')? This is important because, there are two effects that make actual MHC–peptide binding deviate from the IBS ideal. First, individual peptide residue interactions with MHC subtly alter the local environment. These environmental changes affect the structure of neighbouring residues and hence their contribution to the overall binding free energy. Second, the binding free energy is the difference between the free energies of the bound and free states of the peptide. And the free energy of the free peptide could, in principle, be strongly influenced by inter-residue interactions. Both these effects are in direct contravention to the IBS assumption.

Nevertheless, it is possible that at a 'coarse' level of prediction, IBS is a good starting point. We set out at the coarsest possible level of prediction viz. binary prediction: simply predict whether a peptide 'would bind', defined as an IC_{50} below a pre-specified level. In order for any patterns that exist among 'actual binders' to emerge from our procedures, we chose this IC_{50} level close to an antigenically relevant level of 500 nM (Sette et al 1994).

We built an ANN of two layers and with a single output neuron. Out of a database of 463 peptide (9-mer) binding measurements, we varied our training set size over a range and found that a training set size of about 250 measurements was adequate for most purposes. The test set of 151 peptides was kept completely separate and none of the test set measurements was used in training.

In comparison with simple motif searches, the ANN significantly reduced false positives with an average specificity over 90% and positive predictive value of 64%. ANN's sensitivity however was low (45%) at our relevant affinity range (i.e. defining $IC_{50} = 500$ nM as the border between binding and non-binding peptides).

The polynomial method, was complementary to the neural net and had a high sensitivity (85%) at the cost of decreased specificity and positive predictive value (23%).

How can all this help PM?

There are several methods other than the ANN and polynomial methods discussed above for addressing the MHC-peptide binding problem and most of these are familiar to this audience. I hope to prove in the rest of my presentation that these common techniques could prove very valuable in helping bring about PM. For an illustration, consider the case of Alzheimer's Disease (AD) immunotherapy proposed by the Irish company, Elan pharmaceuticals.

The AN-1792 story. Elan's Schenk et al (1999) reported that immunizing PDAPP mice with a 42-amino-acid fragment of β amyloid dramatically reduced the

AD-like pathology characteristic of these mice. Specifically, they reported that immunization of young mice (6 weeks age) almost completely prevented the formation and that of older mice (11 months) dramatically reduced the formation of three pathologies characteristic of AD, viz. β amyloid plaque formation, astrogliosis and neuritic dystrophy. This almost 'picture perfect' story continued to show promise into phase I clinical trials (Schenk et al 2001) of this peptide, dubbed AN-1792.

But early in 2002, the phase 2a clinical trial for the immunotherapy was halted (Check 2002) due to the development of serious nerve inflammation in a minority of subjects. The exact number and proportion of subjects who developed this form of inflammation is unclear as Elan has not yet published the details of their clinical trials. The exact nature of the inflammation and whether any benefit was seen for the rest of the patients are also unknown and this has prompted calls upon Elan to publish their results as soon as possible (Bishop et al 2002).

If the patients who did not develop inflammation showed improvement, then clearly, AN-1792 is a problem waiting for a PM-style solution: is it possible to devise a diagnostic procedure which predicts whether a certain individual will develop the adverse effect?

Since this is immunotherapy, examining the immunological correlates of the occurrence of adverse effects seems like a logical next step. For example, to which HLA molecules do sub-sequences within the vaccine bind? Do these HLA alleles show any relation to the HLA haplotypes of patients who developed inflammation? Given the central role that MHC plays in the immune process, it is possible that Elan is already looking at the HLA types of their patients; however nothing has yet been disclosed.

The susceptibility of specific HLA types for some diseases is well known, such as the celebrated association between HLA-B27 and ankylosing spondilitis. Disease susceptibility is but one phenotype. We can think of drug response as simply another phenotype which might have a correlation to HLA typing. This could particularly be true in the case of immunological pathologies like allergies and autoimmune diseases. In conclusion, I expect that immunological data and consequently immunoinformatics, will have an important role to play in making PM possible.

References

Bishop GM, Robinson SR, Smith MA, Perry G, Atwood CS 2002 Call for Elan to publish Alzheimer's trial details. Nature 416:677

Check E 2002 Nerve inflammation halts trial for Alzheimer's drug. Nature 415:462

Drysdale CM, McGraw DW, Stack CB et al 2000 Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. Proc Natl Acad Sci USA 97:10483–10488

Evans WE, Hon YY, Bomgaars L et al 2001 Preponderance of thiopurine S-methyltransferase deficiency and heterozygosity among patients intolerant to mercaptopurine or azathioprine. J Clin Oncol 19:2293–2301

Gulukota K, DeLisi C 1996 HLA allele selection for designing peptide vaccines. Genet Anal Biomol Eng 13:81–86

Gulukota K, Sidney J, Sette A, DeLisi C 1997 Two complementary methods for predicting peptides binding major histocompatibility complex molecules. J Mol Biol 267:1258–1267

Hall IP 2002 Pharmacogenetics, pharmacogenomics and airway disease. Respir Res 3:10 (available online at *http://respiratory-research.com/content/3/1/10*)

Imanishi T, Akaza T, Kimura A, Tokunaga K, Gojobori T 1993 Allele and haplotype frequencies for HLA and complement loci in various ethnic groups. In: Tsuji K, Aizawa M, Sasazuki T (eds) HLA 1991, Proceedings of the Eleventh International Histocompatibility Workshop and Conference, Yokohama, November 1991, volume 1, Oxford University Press, Tokyo, p 1065–1091

Judson R, Salisbury B, Schneider J, Windemuth A, Stephens JC 2002 How many SNPs does a genome-wide haplotype map require? Pharmacogenomics 3:379–391

Kast WM, Brandt RMP, Sidney J et al 1994 Role of HLA-A motifs in identification of potential CTL epitopes in human papillomavirus type 16 E6 and E7 proteins. J Immunol 152:3904–3912

Krynetski EY, Evans WE 2000 Genetic polymorphism of thiopurine S-methyltransferase: molecular mechanisms and clinical importance. Pharmacology 61:136–146

Lander ES, Linton LM, Birren B et al (International Human Genome Sequencing Consortium) 2001 Initial sequencing and analysis of the human genome. Nature 409:860–921

Parker KC, Bednarek MA, Coligan JE 1994 Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J Immunol 152:163–175

Schenk D, Barbour R, Dunn W et al 1999 Immunization with amyloid-beta attenuates Alzheimer-disease-like pathology in the PDAPP mouse. Nature 400:173–177

Schenk D, Games D, Seubert P 2001 Potential treatment opportunities for Alzheimer's disease through inhibition of secretases and Aβ immunization. J Mol Neurosci 17:259–267

Sette A, Vitiello A, Reherman B et al 1994 The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes. J Immunol 153:5586–5592

DISCUSSION

Rammensee: There are a couple of problems with personalized medicine. These relate not only to the tremendous effort required, but also the principle. For example, if you think of the promiscuous HLA binding type collection of peptides, this essentially involves collecting a few peptides and immunizing a certain group against a disease. Then you run into the risk of that the virus easily adapts because just a very few peptides were taken from it. This counteracts the reason why MHC polymorphism has been established over the last 40 million years.

DeLisi: The key is to use conserved peptides. If you have multiple variants of a virus sequence, you then look for conserved peptides and then use cocktails of these. You choose your cocktails such that there is no known strain that has a mutation in every one of the components of the cocktail. If you do this,

combinatorially there are a lot of possibilities depending on the person's haplotype. You end up with a set of antibodies that are directed against 100 conserved peptides, and you determine which of these bind to which allele and then have reagents against those peptide–allele combinations. When the person comes in you test with a reagent to see what they are presenting, and then use a therapeutic based on this. This is do-able because genome sequencing for viruses is easy. The hard part is that for most viruses we don't have 1000 or 10 000 different strain sequences, so you would need to construct viral mutants.

Rammensee: Personalized medicine should not aim at ethnic groups or other large groups, but individuals.

Gulukota: If you take it to the extreme of personalization, then even for a person I am not suggesting that you give a single peptide, but a cocktail. You need multiple pathways for killing a virus.

DeLisi: The ethnic group aspects are a slightly different context. While there are enormous numbers of MHC alleles, there are probably just 15 alleles that will cover 90% of the world's population. The key here is that we only have to think about some finite number of alleles and a finite number of peptides that bind to those alleles. This is an industrial problem, and we are doing this with a start-up.

Rammensee: But you might not find conserved peptides for all of these 15 alleles.

DeLisi: That's true, but I am thinking of a therapy not a vaccine. It might end up that we can only treat a certain subfraction of the population. That is OK; it is better than nothing. If I can treat 50% of people with HIV I will be very happy.

Margalit: Charles DeLisi, have you looked at the sequences of the viruses and checked whether immunodominant peptides are conserved?

DeLisi: It is purely empirical.

Gulukota: I don't think immunodominant epitopes would be conserved.

Brusic: The immune responses are more complicated than just considering 15 HLA alleles. If we take a group of people who share the same HLA allele and then challenge them with an antigen, they often respond to different peptides.

DeLisi: That's for a vaccine; I'm talking about a therapy, where these problems don't occur.

Borras-Cuesta: If you do therapy you can use almost any helper. But then, if you don't get rid of all the viruses, in the absence of Th memory you will be back to square one. To do this you'd have to try to find a helper peptide from the viral or tumour antigen, and use this helper for immunization experiments in order to generate Th memory specific for the antigen.

DeLisi: That's true for cell based immunotherapies, but there are other options.

Borras-Cuesta: Provided that there are no reservoirs.

DeLisi: Quite. But in this case, there's nothing you can do.

Petrovsky: In terms of applicability and what will be taken up first, I think personalized medicine is more about identifying the individuals who won't react

or who will have adverse reactions than predicting which individuals will respond. In other words, we may have a traditional vaccine and we know that there is a percentage of people who will get encephalitis or other severe reactions. We accept that we may harm one person for every 100 000 people who receive the vaccine and are protected. The anti-vaccine lobby focuses solely on that one person who suffers a side effect. If we could identify that person before they are vaccinated, this would be a useful role for personalized medicine. This is what might come first, rather than developing cocktails of peptides to use as therapy.

Gulukota: I think you are completely right.

Borras-Cuesta: The strategy that Charles DeLisi was suggesting is logical. You know which HLA class I and II the patient has, and you know whether they respond. The logical thing then is to go ahead because you know you are going to induce a response. The only thing you don't know is whether or not you will kill the pathogen. With the alternative approach you have to fish out which people will and won't respond. This requires a tremendous effort. I am not saying it shouldn't be done, but it would necessitate waiting for a long time before implementation of the strategy.

Gulukota: I still think Nikolai Petrovsky is right. There will probably be a correlation between drug response phenotypes and HLA typing or genotyping. This will probably come before than tailor-made medicine for specific individuals. You are right that this is exploratory rather than hypothesis-driven biology, but at least the road is clear. I am sure that someone at Elan or Wyeth is looking at the HLA differences between those who had encephalitis and those who didn't. There has to be some immunological basis for this. It will be easy to find once we look for it. However, if we are going to design drugs specifically for individual haplotypes, this will take longer.

De Groot: I can speak from experience at my company. People doing therapeutics having looked at the EPO (erythropoeitin, a protein used to increase red blood cells) event, where EPO was associated with some adverse events related to antibody responses. Companies are now asking my group to analyse their therapeutic proteins for epitopes. What is really interesting, because this will probably surprise many, is that most of these therapeutic companies have not analysed the HLA of the patients who had the adverse events. The field is new. Then their next question is whether it is possible to reduce the immunogenicity of the molecule. All of us know that this is fraught with problems: if you change one amino acid then you may increase the binding to another allele. Do we have models for all the alleles that exist? No. Industry is getting very involved in this without really understanding what they are embarking on.

Rammensee: Could you say more about the problem with EPO?

De Groot: We have worked on thrombopoietin (TPO), actually, with Genentech.

Rammensee: Is there polymorphism in the protein in different people?

De Groot: I haven't looked at EPO, but we have looked at TPO. The reason I haven't looked at EPO is because there are no HLA data.

Rammensee: Is there polymorphism in TPO? If there is, then one should personalize the recombinant protein.

De Groot: I don't know that, but there are regions in TPO that contain promiscuous epitopes. First you have to show an association between the HLA type and the adverse event, which has not been done. Then you have to show that if you change the molecule you get less of the adverse event, which has also not been done.

Marsh: I don't find it surprising that companies haven't analysed the HLA data. Therapeutic companies are waking up to this idea of HLA, and there is an awful lot of HLA typing that is currently being done.

De Groot: That is helpful for us because we will then start understanding whether there is an association between the presence of an epitope within a 'self protein' such as EPO or TPO, and whether the recognition of that epitope is related to the occurrence of adverse events, or side effects, from therapy in some patients.

Silva: What are the costs of this sort of treatment? If you have to test every single HLA group then the cost will be higher.

Gulukota: When you look at cost of therapy, you have to look at the overall cost. It is fashionable to complain that drugs are expensive, but if you look at the difference in the amount of money that is spent before and after a drug has been released to the market, invariably the amount of money spent after is lower. This is because before the drug therapy the alternatives were plain suffering, failed drug treatments or surgery, for instance. The overall cost is almost always lower.

De Groot: What about FDA concerns and making GMP quantities of these peptide pools? Isn't that a complication? Would you have to get each patient's personalized lot of peptides approved separately in preparation for a phase I trial?

Rammensee: We should first worry about what makes sense rather than the FDA policy. The FDA might then see sense. It would not be possible to test personalized drugs in 100 people.

Gulukota: In personalized medicine we are talking about diagnostics, which have a less stringent standard than new drugs.

Rammensee: Regarding the costs for the genome data which one needs for this sort of work, gene chips will become much cheaper in the future because there is a lot of competition. Lee Hood suggests that in a few years it will be possible to sequence the entire genome of one person in a single day.

Silva: Will this happen in the third world as well?

Rammensee: This depends on the price. Cheap genotyping testing combined with a certain drug might have a better efficiency than conventional treatments.

Obviously, these technologies will be applied in rich countries first because they are expensive at the beginning.

Borras-Cuesta: Having lived in the third world myself I don't think this is true for the third world. These people just do not have enough money. One dollar in Africa is a fortune. I am not against personalized medicine, but it isn't viable for the third world.

Petrovsky: Are we able to come up with an immune panel that companies doing trials on immune therapies should be using? If they are going to do HLA typing, how deep should the HLA typing be? Should they be looking at a cytokine chip with 200 cytokine genes that can be easily be done? If we are going to do trials ourselves, what depth of testing should we do?

Rammensee: It depends on the disease or condition you are looking at. I don't think we can have a big chip and test for everything in every person.

Petrovsky: But if we are looking at the immune system there must be a general panel. With HLA typing, we wouldn't just HLA type people we are immunizing for hepatitis, because HLA typing almost certainly has relevance for understanding responses to all vaccinations. Hence, HLA typing should be part of a general immune panel. The panel should be a generic process that should always be done when we carry out immune intervention. Similarly, you could argue that there are other things that you would do as a generic panel, and then on top of this you would have other specialized tests that may be relevant to particular diseases or interventions. I think it is an interesting idea to try to think what we should include on a core panel using current technologies.

Gulukota: I would test for HLA types, cytokines and blood groups.

Rammensee: For cytokines you would probably look at RNA expression.

Kellam: It is hard to say that you would pick just cytokines for arrays: would you also therefore pick the cytokine receptors and their respective intracellular signalling molecules, for example? You want to capture lots of data on a diverse system and then work it out empirically or model the functional networks rather than trying to guess the few cytokines that you would want to put down in the first instance.

Petrovsky: The great thing about chips is that whether it has 60 or 60 000 spots, it doesn't matter so much in terms of cost so the more the merrier.

DeLisi: You can develop a catalogue of common alleles of the immune system, for example, and this would involve a finite amount of genome sequencing, of say probably 50 individuals. The immune system has perhaps 5000 genes. This could be done within a couple of years and would give us a full catalogue of two or three common alleles per gene. Once you have that catalogue you can do association analyses with various diseases. I think immune system arrays detecting alleles would be worthwhile.

Littlejohn: Isn't one of the problems actually identifying those 5000 genes?

DeLisi: You could look at expression levels in immune cells, such as subpopulations of lymphocytes.

Petrovsky: Much of that data is already available as I know, for example, of groups that have done hundreds of gene expression array analyses on macrophages and T cells.

Gulukota: The thing is, we need to look only at genes that show genuine polymorphism in the population.

DeLisi: All genes typically have two or three common alleles, except for HLA, which has many more. You can determine every one of these simply by full genome sequencing. Within the next couple of years this will be routine. You can get a whole catalogue of all common alleles. The estimate is that just 50 genomes will be needed for this.

Brusic: Some of these assumptions are fine, but genetic factors include many tissue-specific elements, particularly when we consider cytokines and various receptors. There are factors such as promoter elements which are general and there are factors which are specific to an individual tissue for the same gene.

Lefranc: We also need to take into account the antigen receptor specificity. This will be an added level of complexity. Sequences of the variable domains of immunoglobulin or T cell receptors with known specificities, particularly sequences of the V–J and V–D–J junctions of the T cell receptor chains need to be added. It means a lot of work but the experimental methodologies to obtain the data are available, and there are IMGT software tools (IMGT/V-QUEST, IMGT/JunctionAnalysis) to analyse them.

Rammensee: This would require that we first know the antigen epitope exactly, and then that we take the T cell, and do a gene profile of that specific T cell.

Lefranc: This morning we have been talking about the peptide and the MHC, but the third component is the T cell receptor (TCR). The three have to go together. It is not too difficult to include the recognition by the T cell in our approach. We indeed need to see how the peptide presented by the MHC is specifically recognized by the TCR.

Perelson: If I gave you an immunogenic protein and I asked you to tell me within any individual here what sort of T cell would respond to that, how would you do this?

Lefranc: The problem is much more fundamental than this. We want to know which kind of peptide binds to the MHC and which kind of complex is recognized by T cells. We need *in vitro* tests to say that a peptide linked to a given MHC can activate a specific T cell. We now have all these kinds of approaches with HLA tetramers and so on. Things are going in the right direction and the methodology is coming. The MHC tetramer is a good way to catch the specificity of the TCR.

Perelson: We are still a long way from having MHC tetramers for every MHC type.

Lefranc: We are discussing which directions we want to go in. I am suggesting we need to keep in mind the peptide–MHC–TCR interaction. It is a complex story.

Borras-Cuesta: What you are saying is important. The final outcome here is recognition by the T cell. I would like to draw everyone's attention to something I have been thinking about for a while. When you have, for example, a peptide that is recognized by HLA-A2, you more or less know the motifs and which amino acids point to the TCR. What should be the characteristics of residues that point to the TCR and become well recognized? They would be the amino acids that have a high tendency of interacting with other proteins. For example, Singh and Thornton have published a table which is a 20×20 matrix of side-chain interactions in proteins (Singh & Thornton 1992) According to this table, one of the amino acids that gives a better interaction is tryptophan, and another is histidine. Tyrosine, aspartic and glutamic acid are also important. On this basis you could predict whether one epitope would be well recognized by a TCR. But this is good only for viruses and not tumours. In tumours there has been clonal deletion, and in clonal deletion the easily recognized ones are eliminated; the only ones you can use are the intermediate or mediocre ones. I agree, binding to MHC is only part of the story. Binding and recognition are both important.

Gulukota: Also, when we are talking about immunoinformatics, what we might need to start thinking about is that if we want to put together a simple database of immune data, what would go in that? MHC–peptide binding and T cell recognition epitopes are obvious candidates, but we also need phenotypic data such as HLA disease associations and HLA drug response associations. Analysis of all these data to give personalised medicine is one goal that we can work towards.

Schönbach: We also need SNPs and epigenetic data, particularly if we want to design a therapeutic vaccine against cancer.

Reference

Singh J, Thornton JM 1992 Atlas of protein side-chain interactions, Vols I & II. IRL Press, Oxford

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

From immunome to vaccine: epitope mapping and vaccine design tools

Anne S. De Groot*† and William Martin†

**TB|HIV Research Laboratory, Brown University, International Health Institute, Box GB473, Providence, RI 02912, and †EpiVax Inc, 16 Bassett Street, Providence RI 02903, USA*

Abstract. Since the publication of the complete genome of a pathogenic bacterium in 1995, more than 50 bacterial pathogens have been sequenced and at least 120 additional projects are currently underway. Faced with the expanding volume of information now available from genome databases, vaccinologists are turning to epitope mapping tools to screen vaccine candidates. Bioinformatics tools such as EpiMatrix and Conservatrix, which search for unique or multi-HLA-restricted (promiscuous) T cell epitopes and can find epitopes that are conserved across variant strains of the same pathogen, have accelerated the process of epitope mapping. Additional tools for screening epitopes for similarity to 'self' (BlastiMer) and for assembling putative epitopes into strings if they overlap (EpiAssembler) have been developed at EpiVax. Tools that map proteasome cleavage sites are available on the Internet. When used together, these bioinformatics tools offer a significant advantage over traditional methods of vaccine design since high throughput screening and design is performed *in silico*, followed by confirmatory studies *in vitro*. These new tools are being used to develop novel vaccines and therapeutics for the prevention and treatment of infectious diseases such as HIV, hepatitis C, tuberculosis, and some cancers. More recent applications of the tools involve deriving novel vaccine candidates directly from whole genomes, an approach that has been named 'genome to vaccine'.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 57–76

New tools emerging from the informatics revolution are likely to have a dramatic impact on vaccines, accelerating the development of new vaccines, enabling the re-engineering of existing ones, and overcoming traditional barriers to vaccine design. These new tools are urgently needed, since effective vaccines against two pathogens that are responsible for global epidemics — HIV and tuberculosis (TB) — have yet to be successfully developed. Despite years of effort, only one HIV vaccine is in phase I trials, and new vaccines against TB are only now entering the clinical trial pathway. Why the delay? How can new tools developed

in the sphere of immunoinformatics accelerate the process of TB and HIV vaccine development?

The development of vaccines against HIV and *Mycobacterium tuberculosis* (Mtb) has proven more difficult because the correlates of immunity to the pathogen have yet to be well-defined and humoral response to a single protein or set of proteins has not been sufficient to provide protection. Protection against HIV and TB appears to be linked to cellular immune responses (by T helper cells and cytotoxic T cells) to a diverse set of proteins. Vaccines that effectively generate cell-mediated response are needed to provide protection against these pathogens (Seder & Hill 2000).

Selecting the correct antigen or antigens has also been a stumbling block for vaccine development. New genome analysis tools such as microarrays, bioinformatics, immunoinformatics, and high-throughput immunology assays are enhancing our ability to derive proteins or antigens of interest from the genomes of pathogens and are contributing to the development of new concepts in vaccine design such as 'multi-epitope' or 'epitope-driven' vaccines. These tools have also allowed scientists to better define the 'Immunome', that is, the set of information derived from a pathogen that stimulates an immune response.

New vaccine delivery approaches have also been introduced in the last two decades. These include the use of bacterial and viral vectors such as *Salmonella* (Lowe et al 1999), *Listeria* (Lieberman & Frankel 2002), vaccinia and other poxviruses (Stephenson 2001), and adenovirus (Sharpe et al 2002), the use of 'naked DNA' as a means to deliver vaccine components (Johnston & Barry 1997) and the development of new vaccine delivery tools such as gene guns. This article addresses the development of new concepts, tools, and approaches that may accelerate vaccine development from genomic information.

Defining the immunome

In general, host immune response to a pathogen is thought to be due to a number of pathogen-specific responses (provided by antibodies; T helper cells, which drive antibody response; and CTL, for intracellular pathogens). The T cell response is stimulated by the presence of short peptides or epitopes, that are derived from pathogen-specific antigens by antigen presenting cells and presented to T cells in the context of MHC surface proteins (major histocompatability complex molecules, or MHC). Whether the immune response is directed against a single immunodominant epitope or against many epitopes, the generation of a protective immune response does not appear to require the development of T and B cell memory to every possible peptide from every antigen in the entire pathogen. T and B cell responses to the ensemble of epitopes derived from

selected antigens (and not to the whole pathogen) appear sufficient to provide protective immunity.

Consider for example the hepatitis B virus (HBV) vaccine, the cowpox virus, known as vaccinia, which is used prevent smallpox infection, and BCG vaccine, which is used to prevent TB disease. The HBV vaccine consists of a single recombinant protein, separated from the other proteins of HBV. Antibodies developed in response to this protein-based vaccine protect against hepatitis B infection. Thus only HBV protein, and not the entire virus, is needed to generate a protective immune response. While the smallpox and TB and their vaccines are related, they are not identical. Presumably the protective immune response against the pathogen that is generated by immunization with vaccinia is due to B and T cell epitopes that are conserved between the pathogen and its vaccine. Therefore, vaccines that contain a single protein (HBV vaccine) or a subset of proteins (vaccinia and BCG), or even just epitopes derived from those proteins, may be able to create an immune response to challenge the pathogen that is just as effective as vaccines containing whole proteins or whole pathogens. The set of epitopes, which define the 'immunome' of the pathogen, can be defined and discovered by comparing genome sequences and applying new immunoinformatics tools (Fig. 1).

Comparing genomes: a new approach to vaccine development

The publication of the *Haemophilus influenzae* genome in 1995 (Fleischmann et al 1995) was rapidly followed by the genome of *Mycoplasma genitalium*, one of the smallest free-living organisms (only 470 predicted coding regions; Fraser et al 1995). Using these two genomes as a departure point, the research teams discovered the minimal set of genes necessary for independent survival (those contained in the smaller *M. genitalium* genome). Additional contrasts between the genomes of *H. influenzae* and *Escherichia coli* soon followed (Tatusov et al 1996). These genome-comparison approaches set a useful pattern for future contrasts between organisms.

Selecting antigens that may be excluded from vaccines

Certain proteins perform routine functions and are therefore often conserved across different species of microbes, a feature that may make them attractive for vaccine development. A vaccine containing these proteins might protect across species. However, T cells responding to epitopes derived from these proteins may have been tolerized if the housekeeping proteins also resemble similar proteins in the host (Grossman & Paul 2001). Inducing a response to these epitopes might even induce autoimmunity, since minor variations in epitopes

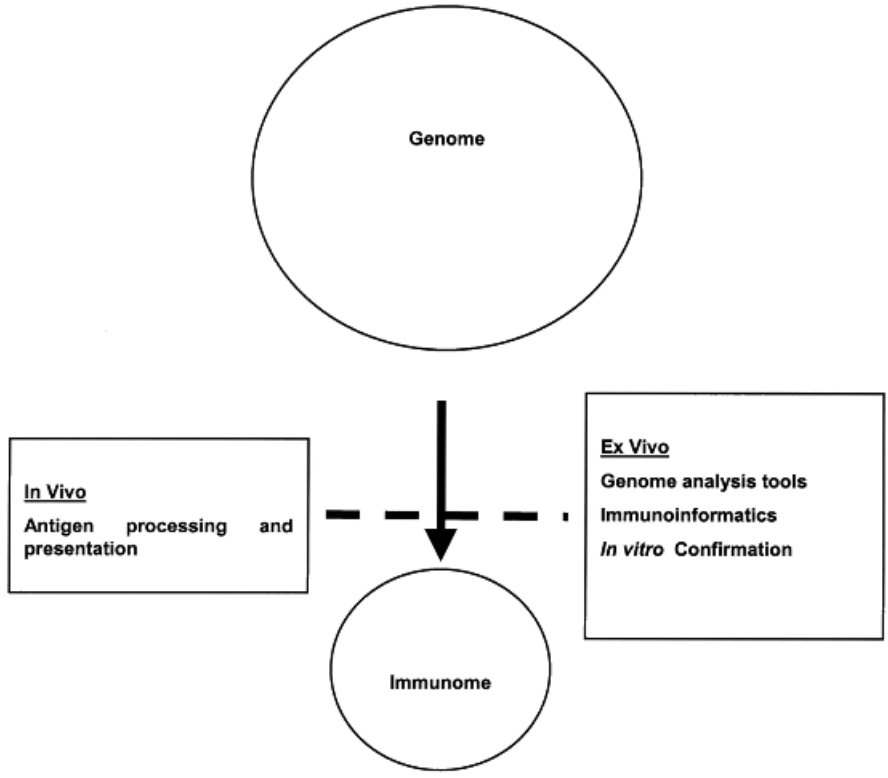


FIG. 1. Defining the immunome.

that are conserved between self proteins and pathogens have been associated with loss of tolerance. In contrast, proteins that are pathogen-specific, particularly those that are secreted by a pathogen, presumably in an attempt to alter the host environment, are more relevant for vaccine development because they are potentially involved in pathogenic activities.

For example, vaccinologists have long been aware of potential differences between the genomes of *Mycobacterium bovis* strain Bacille Calmette Guerin (BCG, the attenuated vaccine used against TB) and *M. tuberculosis*. BCG is not pathogenic in immunocompetent hosts. Genomic analysis of BCG vaccines has now shown that numerous genetic changes (single nucleotide polymorphisms, duplications and deletions) probably occurred during the half-century of ongoing passage of BCG vaccines *in vitro*. A number of genes were also lost. Although the impact of these changes on the protective efficacy of BCG observed in field trials remains to be determined, some researchers have speculated that these genetic deletions have

contributed to making the BCG less effective as a vaccine (Behr et al 2001). A comparison of the genomes of BCG and Mtb, followed by selection of Mtb unique epitopes (not conserved in BCG), is one approach to TB vaccine development currently being explored (De Groot et al 2002).

Selecting antigens that may be included in vaccines

Just as the proteome of an organism can be derived from its genome, the immunome may be derived from comparisons between virulent and avirulent organisms, or between an established vaccine and the pathogen for which it provides protection.

In the case of smallpox, the cell-mediated immune response to vaccinia virus (VV) appears to be one of the major correlates of protection (Erickson & Walker 1993). Some studies have shown that VV-specific, HLA-restricted cytotoxic T lymphocyte (CTL) activity is mediated primarily by $CD8^+$ cells, although low levels of lytic activity by $CD4^+$ cells may also occur. Indeed, the persistent strength of pre-existing cellular immunity to VV has been a concern for vaccinologists who seek to use VV as a vaccine vector: such immunity may interfere with CTL response to VV-vectored vaccine components upon revaccination. Even though the two genomes are quite large and several variants of vaccinia and smallpox have been sequenced, bioinformatics-driven comparisons between the genomes focusing on conserved subsequences may very well lead to the selection of candidates for a novel smallpox vaccine.

Comparisons between the genomes of pathogens and their vaccines may also reveal the immunome if the vaccine is a subunit or attenuated vaccine that represents only a portion of the genome of the original pathogen (Ito et al 2001). This approach may be of use for evaluating other licensed vaccines (Dengue vaccine for example) for expanded use against emerging pathogens (Dengue is also cross-conserved with West Nile Virus). A summary of the types of comparisons described here is provided in Table 1.

Tools

Defining the immunome in silico

Microarray technology is an excellent method for reducing the bewildering array of potential genes to screen in any given genome to a manageable number. Microarrays enable researchers to determine which proteins are expressed during a given phase of the organisms' lifecycle. Comparisons between genomes can also be performed, as can comparisons between genes that are expressed in different 'states' such as those are also up-regulated under 'host conditions' (Schena et al 1995, Cummings & Relman 2000, Dhiman et al 2001). These approaches may be

TABLE 1 Approaches to using genomic information for vaccines

Approach	*Result*
Compare unrelated genomes	Uncover similarities that could be used to make broader vaccines
Contrast genomes of virulent and avirulent strains	Identify antigens associated with virulence for vaccines
Compare genomes of microbial pathogens and their vaccines, or vaccines that might be used	Identify antigens responsible for protective immune response (e.g. BCG and Mtb)

useful for obtaining and collating information relevant to the search for vaccine candidates for a wide range of bacterial and parasitic pathogens.

Bioinformatics tools

One of the major forces driving the development of vaccines from genomes is computational immunology. Computer algorithms for evaluating genomes have evolved in tandem with genome sequencing. We owe the organization of overlapping genome fragments, the derivation of open reading frames (ORFs) encoding putative proteins, and comparisons between newly sequenced ORFs and existing genes to bioinformatics.

Other bioinformatics tools can be used to select genome-derived sequences for characteristics associated with pathogenicity or immunogenicity. For example, protection from disease has, in some cases, been associated with cellular immune response to specific classes of proteins, such as antigens secreted by pathogens into their cellular environment or antigens that span the cellular membrane. Such proteins may now be rapidly identified by scanning a pathogen's genome with computer programs that predict secretory signal peptides (SignalP; Menne et al 2000), transmembrane domains (TMpred; Suhan & Hovde 1998), and lipoprotein attachment sites (Prosite Scan; Falquet et al 2002).

Immunoinformatics tools

Immunoinformatics tools dramatically reduce the time and effort involved in screening potential epitopes (Schafer et al 1998, De Groot et al 2001). Genomes can be scanned and *in vitro* T cell confirmation can be accomplished in a matter of months, instead of years. These methods, coupled with the increased availability of complete and partial genome sequences raises the exciting possibility of building epitope-driven vaccines by directly scanning genomic sequences.

In general, matrix-based T cell epitope mapping algorithms are highly accurate means of searching for putative T cell epitopes. The matrix method enables the assessment of the contribution of secondary anchor residues which engage secondary binding pockets of the MHC molecule. Predictions that examine only the main amino acid anchors have not proven very effective. EpiMatrix (EpiVax, Providence RI), is one of several such matrix-tools (for an extensive review of epitope mapping tools, see De Groot et al 2002). Several *in vitro* studies have confirmed the accuracy of EpiMatrix, in both retrospective (DeGroot et al 1997) and prospective studies (Jin et al 2000, De Groot et al 2001).

Additional approaches to epitope mapping include predictive strategies based on neural networks, threading algorithms, and non-linear functions (Altuvia et al 1995). In several side-by-side comparisons (TB/HIV Research Lab, unpublished comparisons, Yu et al 2002) ANN and matrix-based methods have been found to be essentially equivalent, a finding that provides support for the 'independent side chain contribution hypothesis' on which the matrix methods are based. The most important determinant of the accuracy of the prediction appears to be the actual quality and quantity of the binding data used to derive the predictive method. Following the example of V. Brusic (Brusic et al 1994), researchers who are actively designing epitope-mapping algorithms have amassed large databases of MHC binding peptides.

Recently, the teams of Sturniolo et al (1999) and Zhang et al (1998) proposed that unknown motifs might be predicted by mixing and matching MHC binding pocket characteristics (Fig. 2). Using the approach described by Sturniolo, developers at EpiVax have constructed 74 class II MHC binding prediction matrices (De Groot et al 2003). This new means of developing epitope prediction tools is proving to be extremely useful.

Tools for identifying conserved epitopes

A number of pathogens have been shown to vary between individuals as well as during the course of infection of a single individual. HIV and hepatitis C virus (HCV) are prime examples of such variations; both clades and subtypes (describing variation between infected individuals) and quasispecies (defining variation within a single individual) have been defined. The process of developing vaccines for variable pathogens is complicated by potential variation of key T cell epitopes. However, the Conservatrix algorithm, a bioinformatics tool developed by the TB/HIV Research Lab, can determine which regions are both conserved (across subtypes or quasispecies) and potentially immunogenic. Conservatrix accomplishes this by parsing every sequence in a given database into 9–10 amino acid long text strings. After the algorithm performs a simple string-of-text-based search similar to the approach used by the 'find' function in

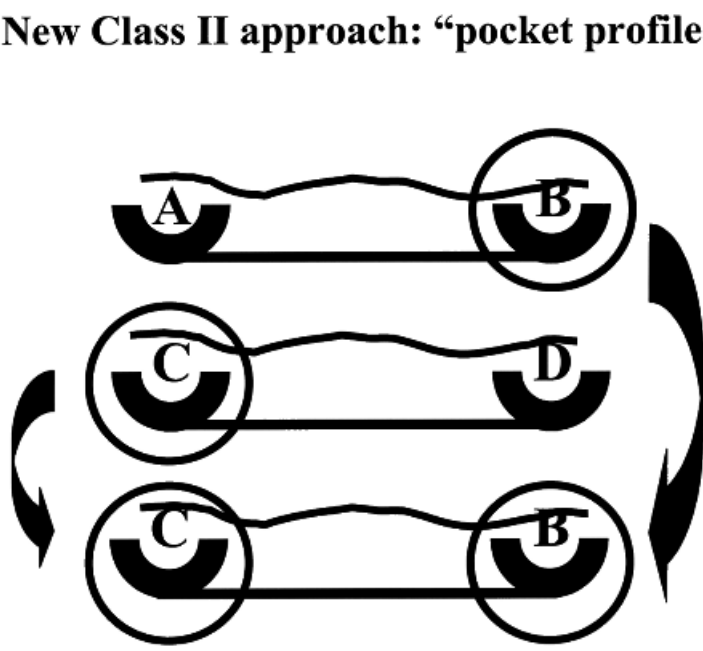


FIG. 2. Pocket profile method.

word-processing programs, each of these text strings is ranked by the number of times it occurs in the set of text strings. Highly conserved peptide text strings are then input into EpiMatrix and ranked for immunogenicity by EBP. This tool has been applied to the analysis of HIV-1, Hepatitis C, and Human Papilloma Virus (De Groot et al 2001 and EpiVax, Providence RI, unpublished results). BlastiMer, another text-based tool developed at EpiVax, compares predicted epitopes for similarity with the human genome. Epitopes that are similar to (less than three amino acids different) or identical to the human genome can be eliminated from the list of epitopes included in a vaccine. Screening vaccine candidates with BlastiMer may eliminate concerns about eliciting autoimmune responses or failure to elicit response due to tolerance of 'self' epitopes.

Epitope-driven vaccines

Epitope-driven vaccines contain only selected sub-sequences, or epitopes, derived from whole proteins. Epitope-driven approaches to evaluating candidate vaccine antigens may even permit a more rapid development of vaccines from pathogenic

genomes since epitopes can be synthesized directly after selection from a genome sequence database. This approach spares researchers labour-intensive steps involving the cloning and expression of the immunogenic proteins prior to development and testing of the vaccine. The concept that an ensemble of epitopes, in the context of the appropriate delivery vehicle, may be able to stimulate a protective response, is driving the development of 'epitope-driven' vaccines in a large number of laboratories (examples include Whitton 1993, Tine et al 1996, Hanke et al 1998, An et al 2000, Morris 2000). Complex vaccines containing T helper and B cell epitopes alongside cytotoxic T lymphocyte (CTL) epitopes derived from a variety of pathogens (such as five viruses and one bacterium) have already been constructed and tested (An & Whitton 1997). A typical epitope-based vaccine construct contains a single start codon with epitopes inserted consecutively in the construct, with or without intervening spacer amino acids. *In vitro* studies of these constructs have confirmed that the epitopes are expressed, stimulate protective immune responses, and do not interfere with one another. Another epitope-driven vaccine approach is to mix several plasmids together, each of which contains genes for different proteins or different minigene epitopes. These vaccines induce no adverse effects, may induce enhanced responses, and may shift responses toward the Th1 phenotype (Tatusov et al 1996). These discoveries suggest that epitope-based vaccines may be particularly useful for pathogens for which no vaccines currently exist.

Epitope strings

EpiVax and the TB/HIV Research Lab have implemented approaches described by other laboratories for enhancing multi-epitope DNA vaccines (Rodriguez & Whitton 2000, Thomson et al 1998). One approach to delivering multiple epitopes in a single plasmid consists of presenting the epitopes as a 'string of beads' without any intervening or 'spacer' sequences separating the individual epitopes (Fig. 3). Several other DNA vaccine researchers have had some success with this approach (An et al 2000). However, in a 'string of beads' construct, the individual epitopes are usually very closely apposed, without their 'natural flanking sequences'—this has raised concern that their proteolytic processing may be compromised, and that peptides other than the specific peptides of interest may be generated as a result of processing (junctional epitopes) (Godkin et al 2001). To address this concern, we have developed the following means of evaluating junctional epitopes: they are paired up, (Fig. 3a) then aligned in sequence, and inserted into a vector plasmid (Fig. 3b). There is some evidence that the introduction of spacer sequences to separate the individual epitopes may help focus the immune response on the specific epitopes (Livingston et al 2001).

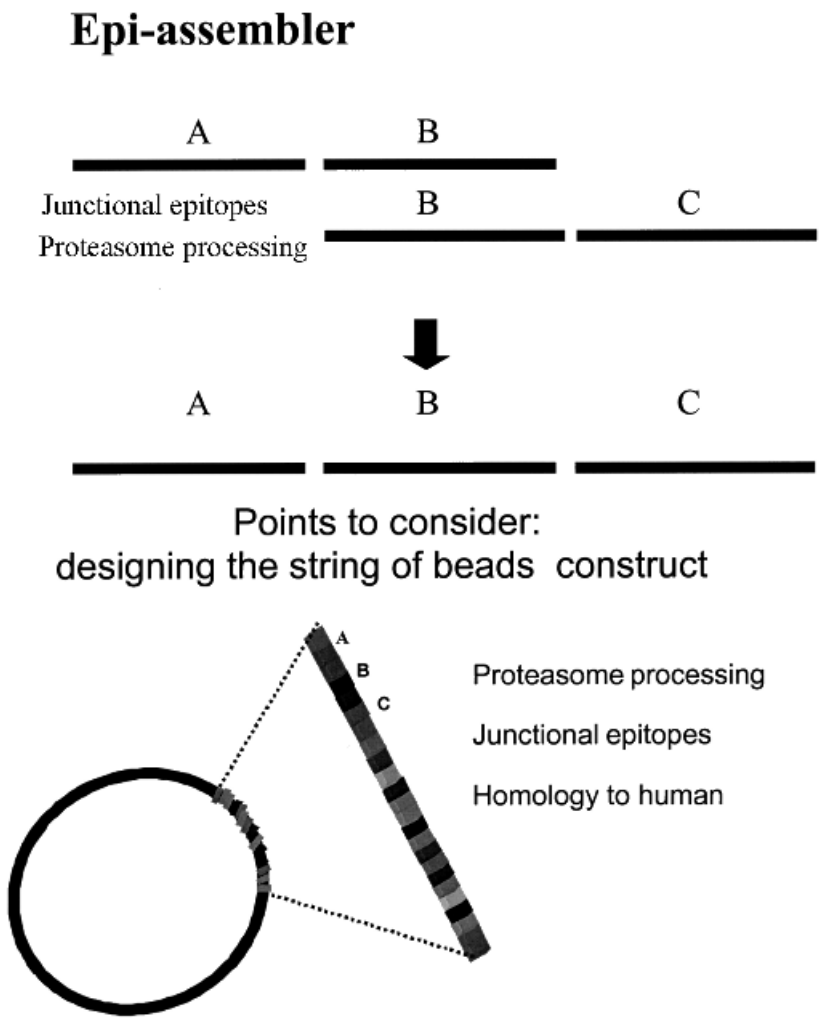


FIG. 3. (a) EpiAssembler. (b) DNA vaccine design.

Spacers and breakers

Studies conducted in murine models have demonstrated that residues flanking an MHC class I epitope strongly influence the delivery of the intact epitope to TAP following proteasome degradation (Holzhutter et al 1999, Thomson et al 1995). However, many of the minigene (multi-epitope) vaccines that have been studied to date have not required the insertion of flanking residues, suggesting that spacers between epitopes are not absolutely necessary to obtain CTL responses. In a recent

comparison of two polyepitope-containing plasmids, Velders et al (2001) have demonstrated the superiority of the construct containing spacers. We have, therefore, implemented the AAY spacer approach described by Kast for our class I epitope-containing prototype vaccines. Published information available on spacers for class II epitopes is relatively limited. Recent work by Livingston et al (2001) with HIV has demonstrated that the use of a standard spacer sequence (-GPGPG-) for HIV vaccine constructs consisting of MHC II-restricted Th cell epitopes disrupts junctional epitopes that would be created by juxtaposing the epitopes and that might compete for degradation or for MHC binding (both G and P are unusual C-terminal anchors for a peptide that binds to class II MHC). This approach has been used for constructs with up to 20 epitopes, in assays where responses were detected to the majority of epitopes.

Directing epitope sequences to class I and II pathways

An additional modification of class II-restricted DNA vaccine epitopes involves the use of signal sequences to target antigenic proteins for display or secretion by infected host antigen presenting cells (APCs). Proteins entering the MHC class II lysosomal degradation pathway do so either via recycling from the cell membrane of the host APCs in which they were made, or more commonly, are shed into the extracellular milieu and taken up by host APCs. The attachment of specific 'signal sequences' to these proteins results in their translation on the membrane-bound ribosomes of the rough endoplasmic reticulum (ER), export into the ER, and subsequent export for either enhanced secretion or membrane localization, which leads to enhanced lysosomal degradation and enhanced activation of the host immune response. For example, the conjugation of a tissue plasminogen activator (tPA) signal sequence to a peptide construct (Malin et al 2000, Li et al 1999) appears to confer the secretory signal necessary for secretion.

Confirming vaccine immunogenicity in transgenic mice

After using *in vitro* T cell assays to select for naturally processed T cell epitopes, it is important to evaluate the ability of vaccines derived from these epitopes to induce an immune response *in vivo*. Non-humanized animal models are not suitable for the evaluation of vaccines designed to induce human HLA-restricted immune responses, as the motifs of epitopes that bind to their MHC molecules often differ from human MHC motifs. Fortunately, a number of transgenic mouse strains that express the most common HLA-A, HLA-B and HLA-DR molecules have been developed (Ishioka et al 1999, Charo et al 2001). A very close correlation has been found between CTL responses in infected individuals and CTL responses induced in immunized HLA transgenic mice (Shirai et al 1995, Le et al 1989, Man

et al 1995). HLA transgenic mice are now routinely used to assay and optimize vaccines in pre-clinical studies.

A number of tools have been developed that incorporate these aspects of vaccine design. One such tool, 'Vaccine-CAD' (computer assisted vaccine design), is under development at EpiVax. This tool incorporates evaluation of junctional epitopes, the insertion of spacers and breakers, requirements for secretion or processing tags, and the evaluation of epitope strings for potential homologies to human gene fragments (Fig. 4).

It should be noted that the concept of epitope-driven vaccines is novel, and only a few of these vaccine constructs have reached the stage of phase III efficacy trials in humans. In the cancer and HIV vaccine fields, where the concept of epitope-driven vaccines is well-accepted, a number of epitope-driven vaccines have successfully passed preclinical tests and are either currently in phase I/II clinical trials or trials are soon to be (Bende & Johnston 2000). Whole protein or attenuated vaccines present well-known risks: the possibility that the live-attenuated vaccine strain may revert to a more virulent form; threats to immunocompromised individuals and individuals with common skin conditions such as eczema (smallpox); and the potential subversion of cellular processes by bacterial and viral proteins to the detriment of the host. Epitope-driven vaccines present several advantages over these other vaccine approaches.

Conclusion

The availability of a large volume of genomic information, coupled with new tools for screening genome sequences *in silico* and refined assays for measuring T cell response to candidate vaccine components have dramatically accelerated the process of vaccine research and development. The discovery of vaccine components no longer appears to depend on understanding the structure and functionality of each of the pathogens' proteins — nor do the proteins have to be isolated or cloned prior to screening. Genome sequences can now serve as a convenient point of departure for *in silico* and *in vitro* approaches to vaccine design.

Acknowledgements and disclosures

Both of the contributing authors are senior officers and majority shareholders at EpiVax, a privately owned vaccine design company located in Providence, RI. These authors acknowledge that there is a potential conflict of interest related to their relationship with EpiVax yet attest that the work contained in this research report is free of any bias that might be associated with the commercial goals of the company.

Initial funding for the TB genome-to-vaccine analysis was provided by a subcontract to the TB/HIV Research Laboratory from an R01 for the sequencing of the 1551 Mtb genome awarded to Dr R. Fleischmann at The Institute for Genomic Research (TIGR). Funding for TB epitope

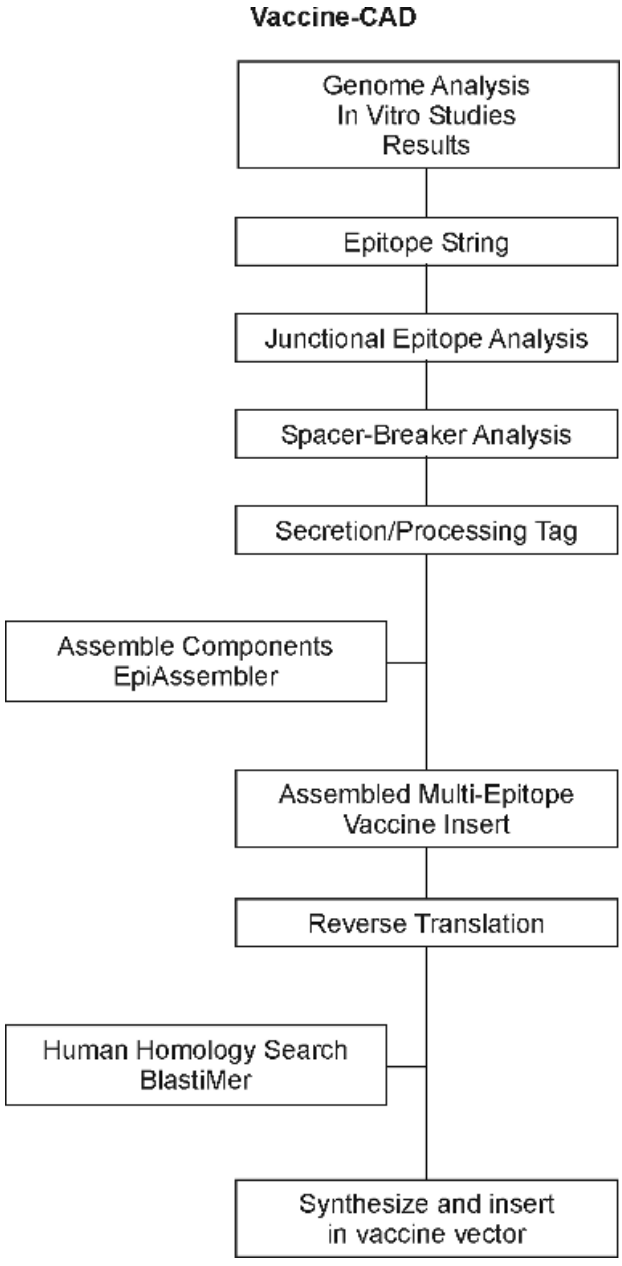


FIG. 4. Vaccine CAD.

analysis and T cell assays described in this manuscript was provided by the Sequella Global TB Foundation in the form of a core scientist award to A. S. De Groot at EpiVax, Inc. Research funding for the HIV studies described in this paper was provided by the Division of AIDS at the NIH through grants to A. S. De Groot (R43 AI 46212, R21 AI 45416, and R01 AI 40888).

References

Altuvia Y, Schueler O, Margalit H 1995 Ranking potential binding peptides to MHC molecules by a computational threading approach. J Mol Biol 249:244–250

An LL, Whitton JL 1997 A multivalent minigene vaccine, containing B cell, cytotoxic T-lymphocyte, and Th epitopes from several microbes, induces appropriate responses in vivo and confers protection against more than one pathogen. J Virology 71: 2292–2302

An LL, Rodriguez F, Harkins S, Zhang J, Whitton JL 2000 Quantitative and qualitative analyses of the immune responses induced by a multivalent minigene DNA vaccine. Vaccine 18:2132–2141

Behr MA 2001 Correlation between BCG genomics and protective efficacy. Scand J Infect Dis 33:249–252

Bende S, Johnston MI 2000 Immunisation. Update: search for an Aids vaccine. AIDS Read 10:526–538

Brusic V, Rudy G, Harrison LC 1994 MHCPEP: a database of MHC-binding peptides. Nucleic Acids Res 22:3663–3665

Charo J, Sundback M, Geluk A, Ottenhoff T, Kiessling R 2001 DNA immunization of HLA transgenic mice with a plasmid expressing mycobacterial heat shock protein 65 results in HLA class I- and II-restricted T cell responses that can be augmented by cytokines. Hum Gene Ther 12:1797–1804

Cummings CA, Relman DA 2000 Using DNA microarrays to study host-microbe interactions. Emerg Infect Dis 6:513–525

De Groot AS, Jesdale BM, Szu E, Schafer JR, Chicz RM, Deocampo G 1997 An interactive website providing major histocompatibility ligand predictions: application to HIV research. AIDS Res Hum Retroviruses 13:529–531

De Groot AS, Bosma A, Chinai M et al 2001 From genome to vaccine: in silico predictions, ex vivo verification. Vaccine 19:4385–4395

De Groot AS, Sbai H, Saint-Aubin C, McMurry JA, Martin W 2002 Immuno-informatics: mining genomes for vaccine components. Immunol Cell Biol 80:255–269

De Groot AS, Rayner J, Martin W 2003 Modelling the immunogenicity of therapeutic proteins using T cell epitope mapping. Dev Biol (Basel) 112:71–80

Dhiman N, Bonilla R, O'Kane DJ, Poland GA 2001 Gene expression microarrays: a 21st century tool for directed vaccine design. Vaccine 20:22–30

Erickson AL, Walker CM 1993 Class I major histocompatibility complex-restricted cytotoxic T cell responses to vaccinia virus in humans. J Gen Virol 74:751–754

Falquet L, Pagni M, Bucher P et al 2002 The PROSITE database, its status in 2002. Nucleic Acids Res 30:235–238

Fleischmann RD, Adams MD, White O et al 1995 Whole genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512

Fraser CM, Gocayne JD, White O et al 1995 The minimal gene complement of Mycoplasma genitalium. Science 270:397–403

Godkin AJ, Smith KJ, Willis A et al 2001 Naturally processed HLA class II peptides reveal highly conserved immunogenic flanking region sequence preferences that reflect antigen processing rather than peptide-MHC interactions. J Immunol 166:6720–6727

Grossman Z, Paul WE 2001 Autoreactivity, dynamic tuning and selectivity. Curr Opin Immunol 13:687–698

Hanke T, Schneider J, Gilbert SC, Hill AVS, McMichael A 1998 DNA multi-CTL epitope vaccines for HIV and Plasmodium falciparum: immunogenicity in mice. Vaccine 16:426–435

Holzhutter HG, Frommel C, Kloetzel PM 1999 A theoretical approach towards the identification of cleavage determining amino acid motifs of the 20S proteasome. J Mol Biol 286:1251–1265

Ishioka GY, Fikes J, Hermanson G et al 1999 Utilization of MHC class I transgenic mice for development of minigene DNA vaccines encoding multiple HLA-restricted CTL epitopes. J Immunol 162:3915–3925

Ito N, Kakemizu M, Ito KA et al 2001 A comparison of complete genome sequences of the attenuated RC-HL strain of rabies virus used for production of animal vaccine in Japan, and the parental Nishigahara strain. Microbiol Immunol 45:51–58

Jin X, Roberts CG, Nixon DF et al 2000 Identification of subdominant cytotoxic T lymphocyte epitopes encoded by autologous HIV type 1 sequences, using dendritic cell stimulation and computer-driven algorithm. AIDS Res Hum Retroviruses 16:67–76

Johnston SA, Barry MA 1997 Genetic to genomic vaccination. Vaccine 15:808–809

Le AX, Bernhard EJ, Holterman MJ et al 1989 Cytotoxic T cell responses in HLA-A2.1 transgenic mice. Recognition of HLA alloantigens and utilization of HLA-A2.1 as a restriction element. J Immunol 142:1366–1371

Li Z, Howard A, Kelley C, Delogu G, Collins F, Morris S 1999 Immunogenicity of DNA vaccines expressing tuberculosis proteins fused to tissue plasminogen activator signal sequences. Infect Immun 67:4780–4786

Lieberman J, Frankel FR 2002 Engineered Listeria monocytogenes as an AIDS vaccine. Vaccine 20:2007–2010

Livingston B, Crimi C, Newman M et al 2002 A rational strategy to design multiepitope immunogens based on multiple Th lymphocyte epitopes. J Immunol 168: 5499–5506

Lowe DC, Savidge TC, Pickard D et al 1999 Characterization of candidate live oral Salmonella typhi vaccine strains harboring defined mutations in aroA, aroC, and htrA. Infect Immun 67:700–707

Malin AS, Huygen K, Content J et al 2000 Vaccinia expression of Mycobacterium tuberculosis-secreted proteins: tissue plasminogen activator signal sequence enhances expression and immunogenicity of M. tuberculosis Ag85. Microbes Infect 2:1677–1685

Man S, Newberg MH, Crotzer VL et al 1995 Definition of a human T cell epitope from influenza A non-structural protein 1 using HLA-A2.1 transgenic mice. Int Immunol 7:597–605

Menne KM, Hermjakob H, Apweiler RA 2000 Comparison of signal sequence prediction methods using a test set of signal peptides. Bioinformatics 16:741–742 (see also SignalP V1.1 World Wide Web Prediction Server, Centre for Biological Sequence Analysis, Department of Biotechnology, The Technical University of Denmark at *http://www.cbs.dtu.dk/services/SignalP*)

Morris S, Kelley C, Howard A, Li Z, Collins F 2000 The immunogenicity of single and combination DNA vaccines against tuberculosis. Vaccine 18:2155–2163

Rodriguez F, Whitton JL 2000 Enhancing DNA immunization. Virology 268:233–238

Schafer JR, Jesdale BM, George JA, Kouttab NM, De Groot AS 1998 Prediction of well-conserved HIV-1 ligands using a Matrix-based algorithm, EpiMatrix. Vaccine 16:1880–1884

Schena M, Shalon D, Davis RW, Brown PO 1995 Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470

Seder RA, Hill AV 2000 Vaccines against intracellular infections requiring cellular immunity. Nature 406:793–798

Sharpe S, Fooks A, Lee J, Hayes K, Clegg C, Cranage M 2002 Single oral immunization with replication deficient recombinant adenovirus elicits long-lived transgene-specific cellular and humoral immune responses. Virology 293:210–216

Shirai M, Arichi T, Nishioka M et al 1995 CTL responses of HLA-A2.1-transgenic mice specific for hepatitis C viral peptides predict epitopes for CTL of humans carrying HLA-A2.1. J Immunol 154:2733–2742

Stephenson JR 2001 Genetically modified viruses: vaccines by design. Curr Pharm Biotechnol 2:47–76

Sturniolo T, Bono E, Ding J et al 1999 Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. Nat Biotech 17:555–561

Suhan ML, Hovde CJ 1998 Disruption of an internal membrane-spanning region in Shiga toxin 1 reduces cytotoxicity. Infect Immun 66:5252–5259 (see also TMpred, European Molecular Biology network at *http://www.ch.embnet.org/software/TMPRED_form.html*)

Tatusov RL, Mushegian AR, Bork P at al 1996 Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with Escherichia coli. Curr Biol 6:279–291

Thomson SA, Khanna R, Gardner J et al 1995 Minimal epitopes expressed in a recombinant polyepitope protein are processed and presented to CD8+ T cells: implications for vaccine design. Proc Natl Acad Sci USA 92:5845–5849

Thomson SA, Burrows SR, Misko IS, Moss DJ, Coupar, BE, Khanna R 1998 Targeting a polyepitope protein incorporating multiple class-II restricted viral epitopes to the secretory/endocytic pathway facilitates immune recognition by CD8+ cytotoxic T lymphocytes: a novel approach to vaccine design. J Virol 72:2246–2252

Tine JA, Lanar DE, Smith DM et al 1996 NYVAC Pf7: a poxvirus-vectored, multiantigen, multistage vaccine candidate for Plasmodium falciparum malaria. Infect Immun 64: 3833–3844

Velders MP, Weijzen S, Eiben GL et al 2001 Defined flanking spacers and enhanced proteolysis is essential for eradication of established tumors by an epitope string DNA vaccine. J Immunol 166:5366–5373

Whitton JL, Sheng N, Oldstone MB, McKee TA 1993 A "string-of-beads" vaccine, comprising linked minigenes, confers protection from lethal-dose virus challenge. J Virol 67:348–352

Yu K, Petrovsky N, Schönbach C, Koh JY, Brusic V 2002 Methods for prediction of peptide binding to MHC molecules: a comparative study. Mol Med 8:137–148

Zhang C, Anderson A, DeLisi C 1998 Structural principles that govern the peptide-binding motifs of class I MHC molecules. J Mol Biol 281:929–947

DISCUSSION

Margalit: Can you explain more about the use of tetramers in this context?

De Groot: We are working on West Nile virus which is a big problem in Israel and now the USA. As a clinician, I know that when I have a patient come into the emergency room with aseptic meningitis, I am now committed to a 5–6 day hospitalization because I have to rule out West Nile. This requires acute serology

and follow-up serology. There is no method of detecting whether this person might have recently been exposed to West Nile when they walk into the emergency room. Our idea is that you might be able to identify T cells in their circulation that are specific for West Nile virus using the tetramer technology. This is what we proposed. It wouldn't be a screen for blood banks because there they just do PCR. In our procedure you could mix the whole blood with the tetramer, run it through the FACS machine in the clinical lab and then detect whether the person had recently been exposed to the West Nile virus. T cell responses are much quicker than the antibody response. We are also very interested in looking at smallpox and vaccinia. You could differentiate people who had been immunized with vaccinia versus people who had been exposed to smallpox in an exposure situation. How else would you know how to quarantine people?

Petrovsky: The problem with that strategy is that it is MHC specific and therefore you can't really develop a generic reagent. It is also epitope specific and you'd need an enormous cocktail to cover most of the epitopes.

De Groot: It is a new idea and there are many potential obstacles. If you include the main five MHC types, you could say for these people with these HLAs, it is not a 100% reliable test in all cases, but if it is positive then you have your answer. A negative result doesn't help.

Rammensee: The critical issue is to verify the predictions. In the case of HIV, this is done with T cells from patients. As I understand it, you test them for recognition of the synthetic peptides, but you don't know whether these peptides are actually processed.

De Groot: We do know that the epitopes were actually processed, because if they are naturally infected and they respond to the peptides, then presumably they have been exposed to a peptide like that which was naturally processed and presented.

Rammensee: There still could be some kind of cross-priming by other cells, and this might not be the same peptide as presented by the infected cell.

De Groot: It could be slightly different, but I don't think it would be very different.

Rammensee: The key question here is whether you could test a cell that is not infected but which is transfected with different genes of interest to see whether they are recognized. This would ensure that these peptides are really relevant.

De Groot: It is hard to know what is in a human. We are going to be doing studies in transgenic mice which are immunized with a construct which contains the whole gene that may contain some of our epitopes. Then we will be coming back (*invitro*) with the epitopes. This may answer your concerns.

Rammensee: Will you have some cells expressing the virus or part of the virus?

De Groot: No, they will be getting a DNA vaccine containing a gene from the virus. We will know if the epitope is naturally processed or not.

Rammensee: It is critical to show that the peptides are naturally processed.

De Groot: This is hard to do in humans who are naturally infected, because we don't know the sequence of the original strain of the virus. If there is a negative response, is it because the epitope wasn't processed, or because the subject have a different strain of HIV and they have never seen that epitope? Presumably, we are finding conserved epitopes, but maybe not. The mouse experiment will be able to answer some of these questions.

Borras-Cuesta: I have a question and a comment with respect to the strategy you have used. You immunize with DNA with multiple epitopes. If I understood you correctly, you said that at the end you had a response when you stimulated with the fuller peptides. What happens to individual peptides? I know from my own experience that sometimes you just get a response to one peptide.

De Groot: It is a very frustrating collaboration. We will be doing individual peptides, but this hasn't been done yet. Currently we only have results showing response to all of the epitopes, in a pool of peptides.

Borras-Cuesta: We have done this in our lab and other people have reported the same thing. If you immunize with these types of plasmids, you can end up having a response to only one of them. It is a crucial issue.

De Groot: This is obviously of great interest to me, because I want to know how many peptides are coming in.

Borras-Cuesta: You mentioned that a peptide was recognized by about 60% of people, and you thought it was a good candidate. In general terms I think you are right, but then you have to ask yourself why these people are still infected. The fact that this peptide is recognised by 60% of the people doesn't make it a good peptide

De Groot: It is big problem. I don't think anyone who is working in HIV knows how to sort this out. There are some data from Bruce Walker looking at the types of T cell responses in patients who are acutely infected. They actually recognize different epitopes in the early infection phase than they recognize during chronic infection. We can't, however, dissect whether this is due to the evolution of the virus or whether it is due to differential processing. The best you can do is base your hypothesis on what is known about effective containment of HIV infection. It appears to be due to the recognition of multiple epitopes. Livingston et al (2001) have shown that if epitopes are presented as single units in a string, this will cause better or broader recognition of a greater number of epitopes than is done if just the gene is presented. You can pack more information in a pseudogene like this than you can using a single gene as an immunogen. It is a mix of hypothesis-driven research and practical experience. We are trying to make something that will work on the basis of what we know. The problem is we don't know whether an immune response to an epitope we identify in this manner is going to work to protect against infection, since the people we are testing already have HIV infection. I should also add that for the Th epitopes we

use long-term non-progressors. Perhaps these are better, because people who do not progress to AIDS do recognize the epitopes.

Borras-Cuesta: So you are hoping to use a good CTL response in vaccination so you would not reach that situation.

De Groot: Yes.

Perelson: I think with HIV vaccines we have to be somewhat careful in distinguishing them from other vaccines. With most vaccines we try to generate what we call sterilizing immunity, so that one person will not become successfully infected by the disease. In HIV no one has been able to establish that state of protection. Most of the vaccines that have been tried in animal studies are non-sterilizing and act as therapeutic vaccines: where we have seen the most success is in generating enough of an immune response to maintain levels of $CD4^+$ T cells that are higher than in unvaccinated animals, and hence extend their lives.

De Groot: This is a major shift in our thinking about vaccines. This deserves emphasis here. In Barcelona, Larry Corey talked about modifying the goal for HIV vaccines. He said that we should not demand that the vaccine act as prophylaxis. Instead, an acceptable new goal is to contain infection. This is a completely new way of thinking about HIV. Perhaps the reason we have 'lowered our expectations' is because we now realize that it will be very difficult to make a vaccine that works prophylactically, so we will accept something that works after infection, by containing the virus better than a non-immunized host. The concept of containing infection is an interesting one, and reflects a shift in the vaccine community in general: therapeutic vaccines will be better accepted in the future on the basis of this work in HIV.

Rammensee: Regarding HIV vaccination, we would probably think in terms of applying the vaccine between phases of HAART.

De Groot: Brigitte Autran in France has set up a network of collaborators and they are looking at therapeutic vaccination for HIV infected patients. The idea is to treat with HAART and get the virus load low, and then vaccinate during HAART with the intention of educating (priming) the maximum number of T cells to respond so that when you take the HAART away you can look at the slope of viral load increase to see whether or not the vaccine is working.

Brusic: Coming back to immunoinformatics, you can see here that there are a number of analytical steps or models that have been put together, starting from genomic information all the way to constructing candidates for vaccines. We would expect that with so many steps involved, errors would creep in and we might not successfully find vaccine targets. Fortunately, the results are quite encouraging. A significant number of peptidic vaccine studies started with predicted targets which were subsequently shown to be functional in patients. This is an illustration of how immunoinformatics can help move the whole field forward. However these are only preliminary studies and we can improve the

synergy between predictions and experimentation. I am pleased to see that you built 74 models for MHC class II peptide binding. This takes a lot of effort. How many people do you have to maintain your prediction system?

De Groot: Just one.

Brusic: We have recently developed a prediction system where a single model predicts peptide binding to an array of HLA molecules. This was achieved by modelling interactions between peptides and multiple MHC molecules. Our single predictive model can in parallel predict peptide binding to multiple HLA-DR molecules, and another model predicts peptide binding to multiple HLA-A2 molecules. This is an example where computational immunology can help us do things more efficiently. There are two sides to the problem — how to discover better vaccines and also how to improve the research methodology. We should strive to advance both these aspects of our research work.

De Groot: Don't you think that we are also expanding our horizons in immunology? Regardless of how you find the protein (by microarray or by direct sequencing), we are looking at the immunogenicity of proteins that people haven't even been able to isolate. What excites me is this 'immunome' problem: how much information (in terms of epitopes) is required in order to get a host to respond effectively to a pathogen? I also don't think that the information required just involves T cell epitopes; it is also B cell epitopes (which I can't model). I think the question 'how much immune information is required to generate a protective response?' is an interesting question to ask. The more we apply these tools, perhaps the closer we will get to answering it.

Reference

Livingston BD, Newman M, Crimi C, McKinney D, Chesnut R, Sette A 2001 Optimization of epitope processing enhances immunogenicity of multiepitope DNA vaccines. Vaccine 19:4652–5660

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

Insights from MHC-bound peptides

Hanah Margalit and Yael Altuvia

Department of Molecular Genetics and Biotechnology, The Hebrew University Hadassah Medical School, Jerusalem 91120, Israel

Abstract. Cytotoxic T cells recognize short antigenic peptides, the processing products of protein antigens, when they are bound to major histocompatibility complex (MHC) class I molecules. Peptide binding to MHC molecules has been studied extensively in numerous laboratories, providing vast amounts of sequence and structure data that have been used as a rich source for bioinformatic research. MHC-bound peptides and their flanking sequences provide information about the sequence requirements of the different processing stages, in particular, the cleavage by the proteasome and the binding to MHC molecules. Elucidation of these sequence requirements sheds light on the evolutionary forces that have shaped and designed these peptides, and should lead to the development of an integrative predictive algorithm. Remarkably, the peptide sequence and structure data are also valuable for the study of biological questions that are apparently unrelated to cellular immunity, namely, sequence–structure relationship and genome annotation. Here we describe our computational analyses of MHC-bound peptides, applied to all these biological topics.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 77–97

Cytotoxic T cells recognize short peptides, the processing products of protein antigens, presented on the surface of antigen presenting cells in association with MHC class I molecules. Binding to MHC is a prerequisite for any T cell-mediated immune response and therefore has been studied extensively by various experimental approaches, which have attempted to elucidate the sequence and structure features that determine the binding specificity. Crystallographic studies revealed the structures of dozens of MHC molecules with their bound peptides, shedding light on MHC and peptide residues that play critical roles in specific binding (e.g. Madden 1995). In parallel, binding experiments and large-scale sequencing efforts of peptides eluted from MHC molecules provided thousands of MHC-binding peptide sequences, which have been compiled in publicly available databases (Brusic et al 1994, Rammensee et al 1999). Multiple sequence alignment of peptides known to bind to a given MHC molecule was used to reveal the residues that are preferred for binding. In turn, these aligned sequences were used for the development of computer algorithms for prediction of

binding peptides based on sequence data (De Groot et al 1997, Parker et al 1994, Rammensee et al 1999).

Although MHC–peptide binding is the most selective stage in the processing of protein antigens, other antigen processing stages contribute to peptide selection. In particular, it was shown that proteasomal cleavage of the protein antigen into short peptides, and the transport of the degraded products to the endoplasmic reticulum by the transporter associated with antigen processing (TAP) are also sequence dependent and play a role in determining the repertoire of immunodominant peptides (reviewed in Yewdell & Bennink 1999). Since both proteasomal cleavage and TAP transport precede MHC binding, the sequences and/or flanking regions in the source proteins of peptides eluted from MHC molecules carry the information that has been used by the cleavage and transport machineries. Therefore, MHC-bound peptides provide a rich source for computational studies attempting to reveal not only the MHC recognition rules, but also the sequence features that play a role in the other processing stages. Here we describe our computational studies that use sequence and structure information to reveal the sequence requirements for MHC binding (Altuvia et al 1995) and proteasome cleavage (Altuvia & Margalit 2000). We also show how the approaches we have taken can be applied to prediction of MHC-binding peptides (Altuvia et al 1997, Schueler-Furman et al 2000), and to selection of peptides with high cleavage potential.

Surprisingly, due to their special characteristics, MHC-bound peptides may provide additional insight, beyond their immunological connotation. First, as they reside both in their native protein and in the MHC groove, analysis of peptides whose structure was solved in these two different environments can be used to examine sequence–structure relationships (Schueler-Furman et al 2001). Secondly, as peptides eluted from MHC molecules reside in proteins that were expressed in the cell, they provide evidence for gene expression at the protein level and are informative for gene verification. By comparing the sequences of the peptides with translation products of the human genome, gene structure and identity can be studied (Altuvia et al 2001). The last two sections of our manuscript discuss both these topics.

Using structural information for prediction of binding peptides

Many computational studies that attempted to unravel the rules governing peptide binding to MHC, used the sequences of MHC-binding peptides. By aligning the sequences known to bind to a given MHC molecule, favourable residues for binding could be identified along the peptide. Synthesis of this knowledge together with that obtained from crystallographic studies has led to understanding of the basic principles that guide peptide–MHC recognition.

Mainly, it was found that certain peptide residues in anchor positions were highly conserved, and contributed significantly to the binding by their optimal fit to residues in the MHC binding groove (reviewed in Madden 1995 and in Rammensee et al 1997). Subsequently, the wealth of MHC-binding sequences were used to generate matrices of coefficients that reflect the suitability of each of the 20 amino acids at each peptide position to bind to a specific MHC molecule. These matrices serve as the basis of many predictive algorithms that evaluate the compatibility of a peptide to bind to an MHC molecule (De Groot et al 1997, Parker et al 1994, Rammensee et al 1999). Their derivation needs, however, extensive experimental work, attempted to obtain a large number of various binding peptides to a given MHC allele.

The approach that we have developed to study peptide-MHC binding does not rely on binding data and sequence information *per se*, but rather uses structural information and employs computational methods developed in the field of computational structural biology. We have shown that these approaches enable us to decipher favourable peptide residues for allele-specific MHC binding, and can be applied for prediction of good binding peptides based on the protein sequence and the solved structures of peptide–MHC complexes (Altuvia et al 1995).

Structural studies indicated that all MHC class I-binding peptides adopt a similar extended backbone conformation in the MHC groove (Madden 1995). Relying on this structural conservation, we developed a structure-based algorithm for MHC binding prediction, adopting the threading approach used in structural biology for protein structure prediction (e.g. Jones et al 1992). The algorithm uses the backbone coordinates of the known peptide fold in a given MHC molecule as a template upon which the sequences of peptide candidates are threaded. For each peptide position we determine the MHC contact residues based on the crystal structures. The interaction of an amino acid at a certain position with its MHC contact residues is evaluated by pair-wise contact potentials (Betancourt & Thirumalai 1999, Miyazawa & Jernigan 1985). By this procedure we were able to explore the suitability of different amino acids at different positions along the peptide for binding to a specific MHC molecule, and to calculate an estimate for the binding energy of a peptide by summing the energies through all peptide positions. We tested this algorithm for several MHC alleles and showed that it succeeds to distinguish between binding and non-binding peptides, and that there is a correlation between the computed binding scores and experimentally measured binding values. Moreover, the algorithm succeeds in ranking highly known immunogenic peptides within all overlapping same-length peptides spanning their respective protein sequences, further supporting its predictive potential (Altuvia et al 1997, Schueler-Furman et al 2000). Still, for MHC alleles where binding data are available, the sequence-based approaches achieve better

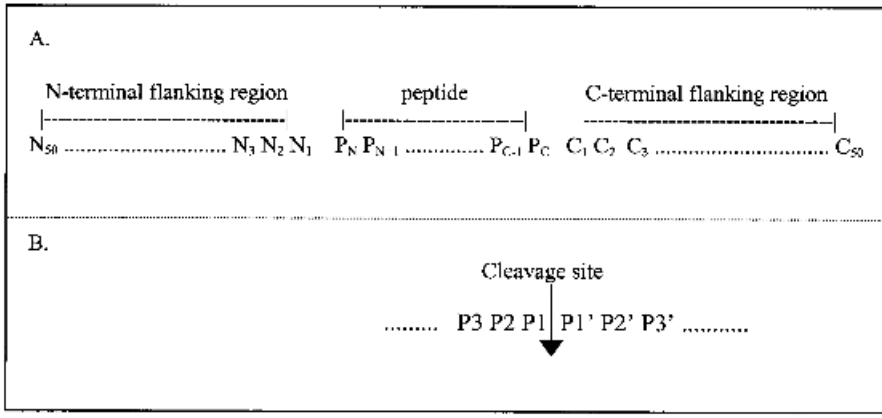


FIG. 1. Terminology used in this study. (A) Nomenclature for positions within the peptide and in its flanking regions. (B) Enzymology nomenclature for a cleavage site. (Reproduced with permission from Altuvia & Margalit 2000.)

prediction performance than the threading approach. Thus, the algorithm is advantageous for MHC alleles that lack binding data but have a solved structure when complexed with peptide, or, alternatively, a structural model of the complex based on known structures.

Extraction of proteasomal cleavage signals

There is accumulating evidence suggesting that proteasomal cleavage is the first step in the processing pathway of most antigenic peptides presented to cytotoxic T cells (reviewed in Koopmann et al 1997, Pamer & Cresswell 1998, Rock & Goldberg 1999). The proteasome is the main protein degradation machine in both the cytosol and nucleus of eukaryotic cells (reviewed in Baumeister et al 1998). Although the structure of the 20 S catalytic core of this multimeric proteinase has been solved and the biochemical activities of the catalytic sites have been characterized, the exact cleavage mechanism and the cleavage specificities are not fully understood.

In a search for potential proteasomal cleavage signals, we performed a rigorous analysis of the residues at the termini and flanking regions spanning 50 residues at both sides of individually sequenced peptides eluted from MHC class I molecules (Altuvia & Margalit 2000, for nomenclature see Fig. 1). We found that the amino acid frequency distributions at the peptides' terminal positions (P_N and P_C), as well as the first position flanking the C-terminus of the peptide (C_1), deviated

significantly from random (Fig. 2). At position P_N basic amino acids and tyrosine were favourable, small amino acids were frequent but not to a statistical significance, and cysteine, proline and leucine were unfavourable. At position P_C hydrophobic and basic residues were favoured. At position C_1 small and basic amino acids were over-represented, while acidic amino acids and phenylalanine were under-represented.

The lack of significant deviation from random at the N-terminal flanking residue (Fig. 2) is not surprising, as it is known that the peptides can be cleaved with an N-terminal extension that is trimmed in later stages (e.g. Craiu et al 1997). As for the C-terminus, experimental studies have suggested that the C-terminus of most cytosolic antigenic peptides is generated directly by the proteasome and is not cleaved further in later processing stages (York et al 1999). Therefore, it is most likely that the amino acid frequency distributions at positions P_C and C_1 describe very closely the amino acid preferences for proteasomal cleavage at the P1 and P1′ positions of a cleavage site, respectively. Indeed, the preferences found at position P_C reconfirm the well-established preference for hydrophobic and basic amino acids at position P1 of proteasomal degradation products, as well as at the C-termini of antigenic peptides. The preferences found at position C_1 reinforce and extend the preferences that were found experimentally at the P1′ position of proteasomal degradation products. Interestingly, a similar preference has been also observed at the N-terminal position of peptides transported by TAP (Daniel et al 1998, Uebel et al 1997, Uebel & Tampe 1999, van Endert et al 1995), a position which coincides with the P1′ position of the peptide's N-terminal proteasome cleavage site (Altuvia & Margalit 2000). The latter observation accentuates the possible role of the signal at position P1′ in determining the cleavage specificity.

We suggest that the amino acid frequency distribution at positions P_C and C_1 can be used to develop a measure for evaluating the cleavage potential between any two residues along a given sequence (Fig. 2 and *http://bioinfo.md.huji.ac.il/marg/cleavage/*). Indeed, in a recent study we used this measure to explore the cleavage potential between internal positions of nonameric antigenic peptides, and found a significantly lower cleavage score between the forth and fifth positions of the peptides. Analysis of the amino acid distribution at those positions revealed that phenylalanine, isoleucine, leucine, methionine and tyrosine, that are favourable at the P1 position of a cleavage site, are under-represented at the fourth position of the peptides (Fig. 3). Correspondingly, proline, glycine, and aspartic and glutamic acid residues, which are unfavourable at the P1 position are abundant at the fourth position. It seems therefore that the amino acid preferences at this peptide position are the mirror image of those at position P1 of cleavage sites. A similar pattern, although somewhat less prominent, is observed at the fifth peptide position. Favourable amino acid residues at position P1′ of a cleavage site, such as alanine, arginine and serine, are under-represented at the fifth peptide position, whereas less

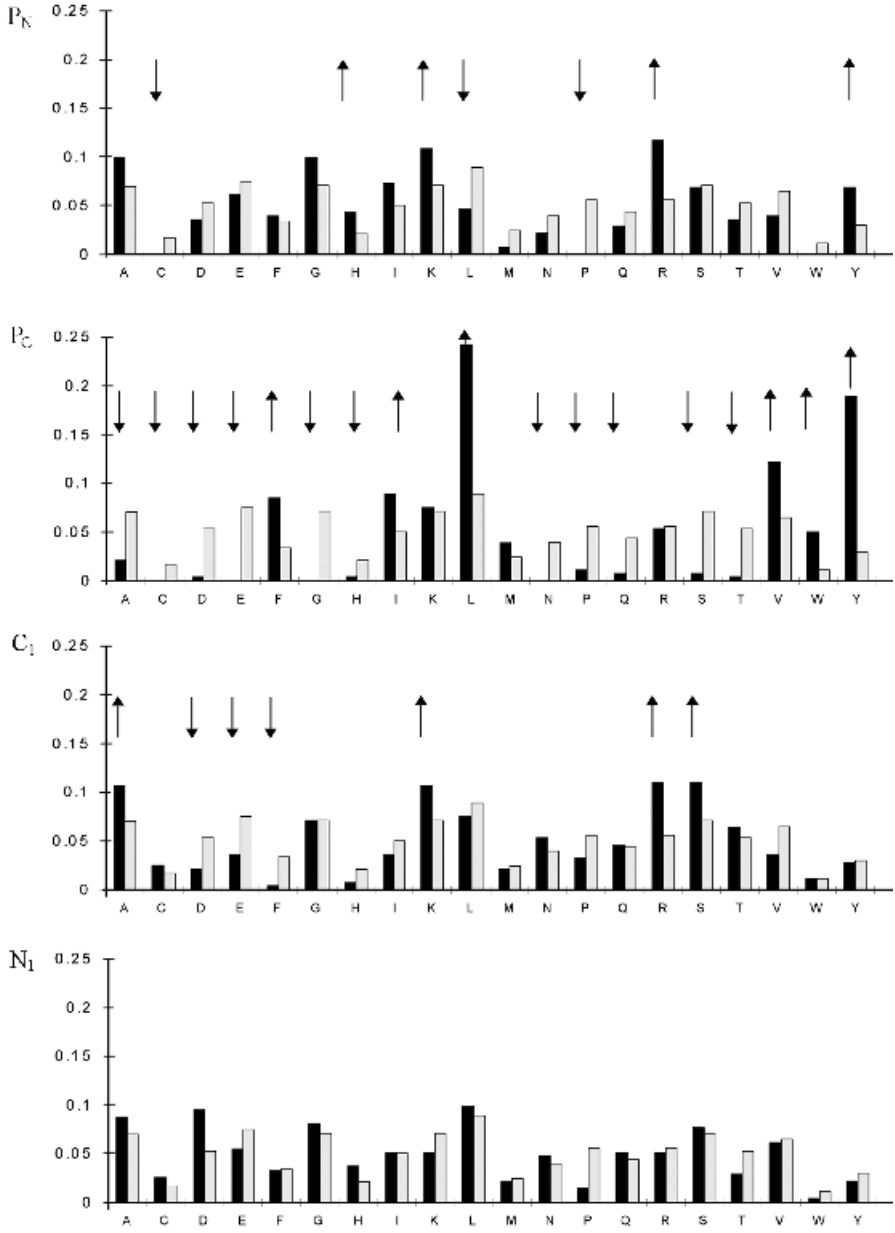
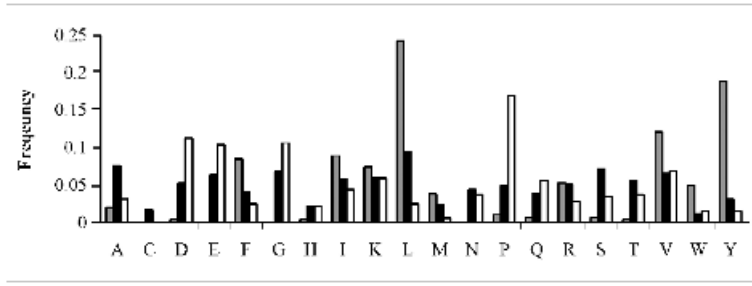


FIG. 2. Amino acid frequencies at positions P_N, P_C, C_1 and N_1 compared to the background frequencies. Dark and light columns represent the observed and expected amino acid frequencies, respectively. Arrows mark statistically significant deviations between observed and expected frequencies (↑, observed > expected; ↓, observed < expected). Amino acids are denoted by their one letter code. (Reproduced with permission from Altuvia & Margalit 2000.)

a.



b.

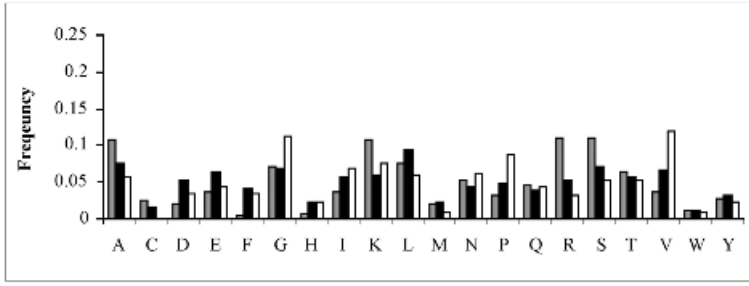


FIG. 3. Amino acid distribution at the central peptide positions. The frequency distribution for each amino acid at the fourth and fifth positions is compared to the background distribution and to the corresponding value at the cleavage sites of the proteasome. The background frequencies were calculated from the database of protein sequences, SWISSPROT. Amino acid frequencies for the P1 and P1′ positions were based on our analysis (Altuvia & Margalit 2000). Peptides in which the central positions are important for the specific binding to the MHC molecule were excluded from the analysis to prevent possible bias. The amino acid distributions at both positions deviated significantly from the background ($\chi^2=170$, $P=0.001$, $n=299$ for the fourth position and $\chi^2=63$, $P=0.001$, $n=278$ for the fifth position). (a) Amino acid frequencies at the fourth peptide position (white bars) as compared to position P1 of a cleavage site (grey bars) and to the background (black bars). (b) Amino acid frequencies at the fifth peptide position (white bars) as compared to position P1′ of the cleavage site (grey bars) and to the background (black bars).

favourable amino acid residues, such as proline, valine and isoleucine are relatively abundant. Taken together, these observations suggest the possible role of the amino acid residues at the central peptide positions in the prevention of internal cleavage. Notably, one of the most frequent amino acids at both positions is proline. Indeed, proline was shown experimentally to prevent internal proteasomal cleavage of peptides (Shimbara et al 1998), and was found to be abundant at position P4 of proteasomal cleavage sites (Nussbaum et al 1998).

It was suggested that proteasomal cleavage plays a role in the selection of immunodominant peptides (e.g. Yewdell & Bennink 1999). If our findings were correct we would expect the residues at position C_1 and the central residues to affect immunogenicity (in addition to the well-known effect of the C-terminus). Indeed, it was recently shown experimentally that the type of residue at position C_1 affects immunogenicity (Livingston et al 2001), and the favourable residues correlated well with those identified computationally (Fig. 2). It was also shown that cleavage within the peptides could reduce or eliminate completely their presentation (reviewed in Niedermann et al 1999, Yewdell & Bennink 1999, York et al 1999). Based on our findings we have developed a quantitative measure that takes into account the two important cleavage considerations of antigenic peptides, namely, exact cleavage at the C-terminus and resistance to cleavage at the centre of the peptides. This measure, which is a linear combination of the internal and terminal cleavage scores, succeeds fairly well in distinguishing between immunodominant and cryptic peptides (our unpublished results).

Structural properties of MHC-binding peptides

MHC-binding peptides are very intriguing for analyses of sequence–structure relationships, as they actually exist in two different structural environments: their native source proteins and the MHC binding groove. We studied the structural properties of the peptides within their two structural environments addressing both the immunological question, regarding possible structural constraints imposed on T cell antigenic peptides, as well as the general question regarding the relationship between sequence and structure (Schueler-Furman et al 2001).

Our study involved 14 peptides that were solved crystallographically both within their native protein and when bound to the MHC molecule (both class I and class II-bound peptides, the latter being recognized by helper T cells). Comparison of their conformations in both environments revealed that while both MHC class I and class II binding peptides showed conserved extended structures in the MHC binding groove, they displayed a large variety of secondary structure types within their native proteins (Fig. 4). These secondary structures ranged from helical through extended to coil. While the native structure of a few peptides was extended and resembled their MHC-bound structure, most peptides adopted entirely different structures in their native conformation. The variety of secondary structure types exhibited by antigenic peptides within their native proteins implies that they adopt the required conformation within the MHC groove independent of their structural background.

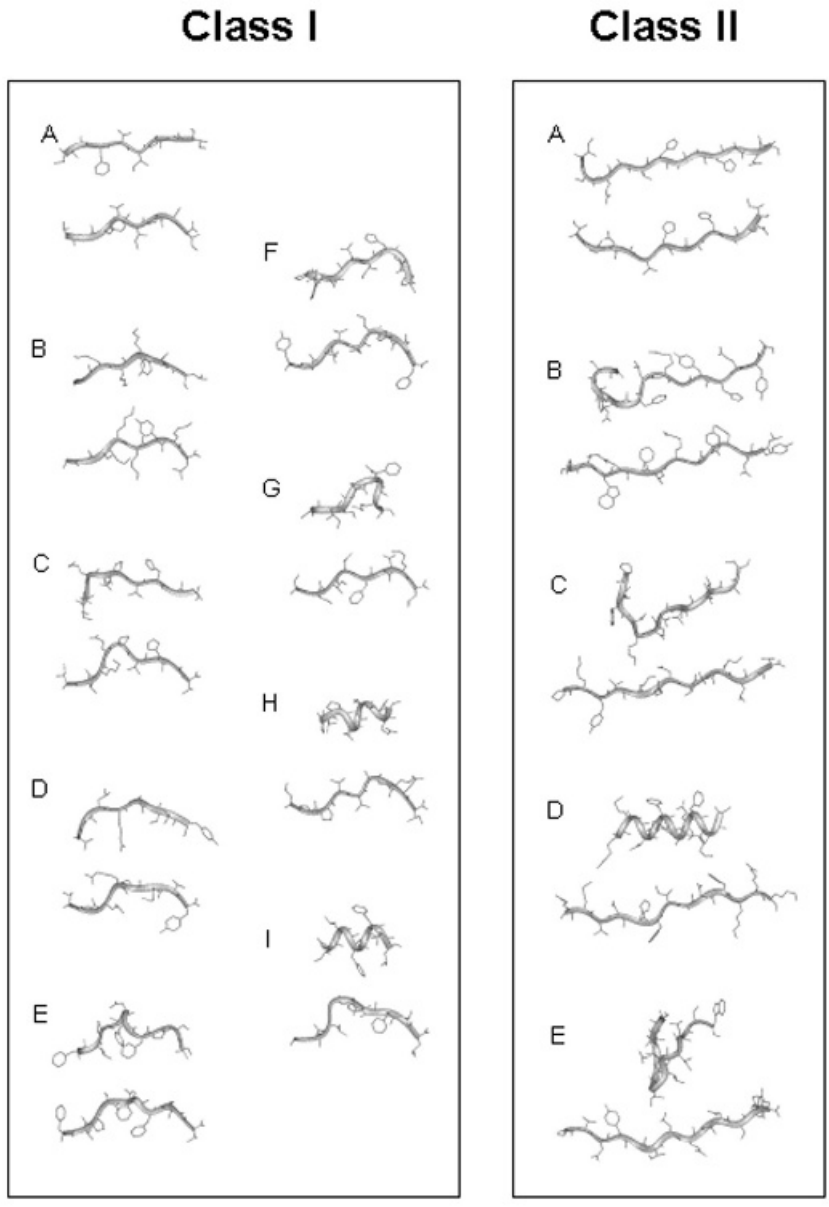
To test this further we organized a database of 67 nonameric class I MHC-binding peptides whose native proteins were solved crystallographically. Ideally, we would have compared the structures of these peptides to their structures within the MHC groove, but the latter have not been solved yet. Still, based on the high structural conservation of MHC-bound peptides, it is conceivable that they adopt similar conformations upon binding. Therefore, their overall structural properties could be compared to those of a group of 21 other nonamers whose structures were solved within the MHC groove. The distributions of three structural properties were examined within each group: secondary structure, C_α–C_α distance of peptide termini, and root-mean-square deviation ($RMS_{C\alpha}$) between all possible pairs of peptides. For all three parameters the native structures were significantly more variable compared to the MHC-bound structures. The secondary structures, and the C_α–C_α distances of peptide termini were also compared to a dataset of random nonameric peptides derived from the same source proteins. No significant difference was found, supporting the notion that the MHC binding potential is independent of the peptide structure in its native protein, and that a peptide can essentially be derived from any region in a protein.

Short subsequences up to nine amino acids long that adopt different conformations when embedded in different proteins were reported before (e.g. Zhou et al 2000) and termed 'chameleons' (Minor & Kim 1996). It has been claimed that the structural environment induces the structure of these short sequences, and that their final structure is influenced by long-range interactions. The MHC-bound peptides expand this collection and support this conjecture. The deep burial of the peptides in the MHC groove justifies their treatment as an integral part of the MHC protein (discussed in Schueler-Furman et al 2001). Thus, they provide additional data of subsequences that adopt different structures in different environments, and set a higher limit of 14 amino acids for the length of such peptides. This has important implications for structure prediction algorithms that are based on structures of short sub-sequences (e.g. Simons et al 1997).

Naturally processed peptides (NPPs) and genome annotation

One of the major roles of the cellular immune system is to destroy cells expressing non-self or mutated proteins. Nevertheless, most of the antigen processing stages are indifferent to the peptide source, and therefore a large fraction of NPPs that are eluted from MHC molecules, originate from self cellular translation products. These include native proteins, as well as 'defective ribosomal products' (DRiPs), (Yewdell et al 2001), consisting of various damaged proteins and protein fragments.

Since the presence of an MHC-bound peptide indicates that its source protein was present in the cell, we suggested that tracing those peptides back to their DNA


Class I
Class II
A
B
C
D
E
F
G
H
I
A
B
C
D
E

source should be very informative for gene annotation and for verification of *bona fide* gene expression. Accordingly, we carried out a comprehensive search comparing hundreds of individually sequenced peptides eluted from MHC molecules from the SYFPEITHI database (Rammensee et al 1999) to all accumulated human sequence data in the form of proteins, mRNA, expressed sequence tags (ESTs), and human protein and mRNA predictions (Altuvia et al 2001). In addition, the availability of the draft of the human genome (*http://genome.ucsc.edu*) allowed us to directly match the peptides against all six translated reading frames of each chromosome.

The detailed search results are available at *http://bioinfo.md.huji.ac.il/marg/IMtoGENE/*. ~73% of the 514 analysed peptides were matched exactly to human proteins and/or translated mRNA (most of these hits were documented previously in the SYFPEITHI database (Rammensee et al 1999). This implies that their genes are translated and expressed at the protein level. This information is especially valuable for verifying the expression of hypothetical proteins derived by conceptual translation. 48 of the above peptides did not match exactly any human chromosome (in all six reading frames). Analysis of those peptides showed that ~63% spanned a splice site and ~14% matched a translated genomic sequence with one nucleotide mismatch. This implies that tracing NPPs can add supporting evidence for putative splice junctions, and can also hint at possible polymorphisms. A small number of peptides did not match any human protein or mRNA sequence, still they matched a human genomic sequence, and/or human EST, and/or other mammalian protein or mRNA sequence. Those hits were especially attractive as pointers to potential new genes.

Summary

MHC-bound peptides provide a rich source of information. Their analysis provides important insight into a broad spectrum of biological questions. Firstly, the unique structural properties of the peptides expand the repertoire of

FIG. 4. MHC-binding peptides as chameleon sequences. Structural pairs of identical peptides in two different structural environments are shown, ordered according to the $RMS_{C\alpha}$ between the peptide structure in the native protein (upper structure) and in the MHC groove (lower structure). Left panel: MHC class I binding peptides. (A) HIV-1 reverse transcriptase (128–135), (B) HIV-1 P17 (24–31), (C) HIV-1 reverse transcriptase (309–317), (D) HIV-1 Nef (75–82), (E) HBV nucleocapsid (18–27), (F) spectrin (190–198). (G) chicken ovalbumin (257–264), (H) HIV-1 integrase (28–36), (I) influenza matrix protein (58–66). Right panel: MHC class II binding peptides (A) chicken ovalbumin (323–334), (B) HLA-A*0201 (128–141), (C) influenza haemagglutinin (306–318), (D) HSP70 (236–248), (E) hen egg lysozyme (50–62). (Reproduced with permission from Schueler-Furman et al 2001.)

subsequences that can fold differently in two unrelated environments. The upper limit of 14 amino acids that we have found might even be set higher with newly solved MHC class II structures with longer peptides. This is of great importance to the general sequence–structure relationship question, and has direct implications for structure prediction algorithms, as it shows the limitations of local homology modelling. Secondly, the 'sampling nature' of these peptides, that constantly present fractions of the content of the cells' proteins, can be used for gene annotation and verification of gene expression at the protein level. Although the analysis described here focused on tracing back human NPPs, we suggest that this method can be a useful annotation tool for other organisms and in particular it can be easily applied to mouse, since the draft of the mouse genome, as well as a large database of mouse NPPs, are already available.

Most importantly, MHC-binding peptides are first and foremost elements of the cellular immune system. They undergo various selective antigen-processing stages, storing in their sequence and structure valuable information of the different cellular mechanisms they encounter. When we piece together these bits of information the remarkable evolutionary compatibility of the different components of the cellular immune system emerges. It has already been shown that the preference for hydrophobic and basic amino acids at the peptides' C-termini is compatible with the requirements for binding to many MHC class I alleles (Rammensee et al 1997, 1999), as well as for TAP binding (Daniel et al 1998, Uebel et al 1997, Uebel & Tampe 1999, van Endert et al 1995) and for proteasomal cleavage (reviewed in Niedermann et al 1999). Likewise, as we have described above, the amino acid preferences at the P1′ positions of proteasomal cleavage sites match the TAP binding requirements at the N-terminus of transported peptides (Altuvia & Margalit 2000, Daniel et al 1998, Uebel et al 1997, Uebel & Tampe 1999, van Endert et al 1995). Also, it seems that there is a good agreement between non-cleavage preferences at the centre of nonameric peptides and the structural and sequence constraints that are imposed by MHC binding. The residues that are unfavourable for cleavage and are abundant at the fourth and/or fifth positions, namely proline, glycine, aspartic and glutamic acids, asparagine and glutamine, are the very same residues that are expected to suit best the structural constraints at the centre of the peptide. Proline and glycine that are known as 'turn formers' (Chou & Fasman 1974) can contribute to the bulge formation at the centre of the peptide, and the hydrophilic amino acids are compatible with the more exposed structural environment that those positions encounter (Madden 1995).

Finally, the bound structure of the peptide enforced by the MHC binding groove seems like a very successful fold, as it is so general and robust as to allow peptides that originate from different structural environments to adopt it. This enlarges the repertoire of potential immunodominant peptides and increases the

chance that in a population there will always be an individual with a haplotype that recognizes at least one peptide within an invader's protein (Schueler-Furman et al 2001).

The 'integrative approach' that evolution has taken in shaping the sequence and structure features of antigenic peptides is impressive. A similar integrative approach is the challenge for future predictive algorithms. The currently available algorithms are based mainly on identifying good binders to MHC molecules. However, binding to MHC *per se* does not guarantee immunodominance, as the peptide has first to go successfully through the preceding processing stages. As shown above, computational analysis of the peptides revealed additional sequence features used by the various processing stages, and the challenge ahead is to incorporate those appropriately in one predictive algorithm.

References

Altuvia Y, Margalit H 2000 Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage mechanism. J Mol Biol 295:879–890

Altuvia Y, Lithwick G, Margalit H 2001 Harnessing the cellular immune system to the gene-prediction cart. Trends Genet 17:732–734

Altuvia Y, Schueler O, Margalit H 1995 Ranking potential binding peptides to MHC molecules by a computational threading approach. J Mol Biol 249:244–250

Altuvia Y, Sette A, Sidney J, Southwood S, Margalit H 1997 A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. Hum Immunol 58:1–11

Baumeister W, Walz J, Zuhl F, Seemuller E 1998 The proteasome: paradigm of a self-compartmentalizing protease. Cell 92:367–380

Betancourt MR, Thirumalai D 1999 Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 8:361–369

Brusic V, Rudy G, Harrison LC 1994 MHCPEP: a database of MHC-binding peptides. Nucleic Acids Res 22:3663–3665

Chou PY, Fasman GD 1974 Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry 13:211–222

Craiu A, Akopian T, Goldberg A, Rock KL 1997 Two distinct proteolytic processes in the generation of a major histocompatibility complex class I-presented peptide. Proc Natl Acad Sci USA 94:10850–10855

Daniel S, Brusic V, Caillat-Zucman S et al 1998 Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. J Immunol 161:617–624

De Groot AS, Jesdale BM, Szu E, Schafer JR, Chicz RM, Deocampo G 1997 An interactive Web site providing major histocompatibility ligand predictions: application to HIV research. AIDS Res Hum Retroviruses 13:529–631

Jones DT, Taylor WR, Thornton JM 1992 A new approach to fold recognition Nature 358: 86–89

Livingston BD, Newman M, Crimi C, McKinney D, Chesnut R, Sette A 2001 Optimization of epitope processing enhances immunogenicity of multiepitope DNA vaccines. Vaccine 19:4652-4660

Koopmann JO, Hammerling GJ, Momburg F 1997 Generation, intracellular transport and loading of peptides associated with MHC class I molecules. Curr Opin Immunol 9:80–88

Madden DR 1995 The three-dimensional structure of peptide-MHC complexes. Annu Rev Immunol 13:587–622

Minor DL Jr, Kim PS 1996 Context-dependent secondary structure formation of a designed protein sequence. Nature 380:730–473

Miyazawa S, Jernigan RL 1985 Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 18:534–552

Niedermann G, Geier E, Lucchiari Hartz M, Hitziger N, Ramsperger A, Eichmann K 1999 The specificity of proteasomes: impact on MHC class I processing and presentation of antigens. Immunol Rev 172:29–48

Nussbaum AK, Dick TP, Keilholz W et al 1998 Cleavage motifs of the yeast 20S proteasome beta subunits deduced from digests of enolase 1. Proc Natl Acad Sci USA. 95:12504–12509

Pamer E, Cresswell P 1998 Mechanisms of MHC class I–restricted antigen processing. Annu Rev Immunol 16:323–358

Parker KC, Bednarek MA, Coligan JE 1994 Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J Immunol 152:163–175

Rammensee H-G, Bachmann J, Stevanović S 1997 MHC ligands and peptide motifs. Landes Bioscience, Austin TX USA

Rammensee H-G, Bachmann J, Emmerich NP, Bachor OA, Stevanović S 1999 SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50:213–219

Rock KL, Goldberg AL 1999 Degradation of cell proteins and the generation of MHC class I-presented peptides. Annu Rev Immunol 17:739–779

Schueler-Furman O, Altuvia Y, Sette A, Margalit H 2000 Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. Protein Sci 9:1838–1846

Schueler-Furman O, Altuvia Y, Margalit H 2001 Examination of possible structural constraints of MHC-binding peptides by assessment of their native structure within their source proteins. Proteins 45:47–54

Shimbara N, Ogawa K, Hidaka Y et al 1998 Contribution of proline residue for efficient production of MHC class I ligands by proteasomes. J Biol Chem. 273:23062–23071

Simons KT, Kooperberg C, Huang E, Baker D 1997 Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 268:209–225

Uebel S, Tampe R 1999 Specificity of the proteasome and the TAP transporter. Curr Opin Immunol 11:203–208

Uebel S, Kraas W, Kienle S, Wiesmuller KH, Jung G, Tampe R 1997 Recognition principle of the TAP transporter disclosed by combinatorial peptide libraries. Proc Natl Acad Sci USA 94:8976–8981

van Endert PM, Riganelli D, Greco G et al 1995 The peptide-binding motif for the human transporter associated with antigen processing. J Exp Med 182:1883–1895

Yewdell JW, Bennink JR 1999 Immunodominance in major histocompatibility complex class I-restricted T lymphocyte responses. Annu Rev Immunol 17:51–88

Yewdell JW, Schubert U, Bennink JR 2001 At the crossroads of cell biology and immunology: DRiPs and other sources of peptide ligands for MHC class I molecules. J Cell Sci 114:845–851

York IA, Goldberg AL, Mo XY, Rock KL 1999 Proteolysis and class I major histocompatibility complex antigen presentation. Immunol Rev 172:49–66

Zhou X, Alber F, Folkers G, Gonnet G, Chelvanayagam G 2000 An analysis of the helix-to-strand transition between peptides with identical sequence. Proteins 41:248–256

DISCUSSION

Rammensee: Your data set is biased in that most of the peptides are HLA-A2 and most peptides are produced by the immunoproteasome, rather than the constitutive one. Most of the sequenced ligands are from cell lines which grow rapidly *in vitro*, and these cells frequently express the immunoproteasome. Only a few of the peptides in your database are from cells known to express constitutive proteins.

Margalit: This is a good thing from my perspective. I was somewhat concerned knowing that I have a mixture of naturally processed peptides that are cleaved by the constitutive proteasome and antigenic peptides that are cleaved by the immunoproteasome. But if you tell me that most of the peptides are cleaved by the immunoproteasome, this is good. We want to discover the cleavage of the immunoproteasome and to integrate it in a predictive scheme.

Rammensee: There is a concern that the middle of the peptide is preferentially selected for cleavage by the proteasome. There are several cases where different peptides from one protein bind to different HLA molecules, and the overlap that occurs to create binding to one HLA molecule requires cleavage of, let's say, an HLA-A2 epitope. Thus, a certain cleavage of the proteasome might create the correct C-terminus for one HLA molecule and at the same time destroy the ligand for another one. Perhaps this finding might be a consequence of the bias for HLA-A2. A larger database might be able to consider such situations.

Kesmir: We have done this. We took Hannah Margalit's database and we knew it was biased towards A2, so we added a lot of peptides from MHCPEP so that these 55 MHC molecules were more or less similarly represented. We ended up with more than 1000 peptides. Indeed, you lose the signal on PN, but the signals on PC and C2 remain.

Rammensee: With MHCPEP we have a new problem that it is not ligands, it is just binding and ligands together. It is a mix of everything.

Kesmir: But at least we got rid of the bias problem!

Stevanović: You have to consider that this cleavage is just one cleavage in the whole protein. If we analyse the peptides that are created in the proteasome — if we digest whole proteins — we see there are 100–300 cleavage sites in the protein and we have no clue whether your cleavage sites are the more dominant ones. It is important that these peptides are created, but this will not give us much information about the total specificity of the immunoproteasomes, because the quantitative effects of the proteasome cutting one protein are completely neglected.

Margalit: I accept this, but in the context of the antigenic peptide I hope it may make some contribution. I'd like to mention that Arieh Admon in Haifa, an expert in mass spectrometry, took cells from lung cancer and isolated the peptides bound

to MHC molecules from these cells. The proteins that were expressed in lung cancer are in a database on the Internet. We took these data and analysed the peptides, and we found very similar preferences at the P1 and P1′ positions with just one exception that didn't fit. We found leucine in the P1′ position. All the other preferences were very similar to the ones we found.

Rammensee: You might have a constitutive proteasome in lung cancer cells. This changes the P1 position.

Margalit: It is the P1′ position.

Petrovsky: What was the MHC that he was eluting from?

Margalit: I don't remember. However, we used these data to test our predictive algorithm. The other predictive algorithms were developed on the degradation products of the constitutive proteasome. When they were applied to MHC binding peptides they succeeded in predicting 64 out of 160. They said that they need a lot more data to refine the predictive algorithms to make better predictions. It is still work in progress. I believe there is still no good prediction algorithm for cleavage of the antigenic peptides.

Brusic: A major problem with this is that we still have a strong bias where we have plenty of data on HLA-A2 or HLA-A1, but less data on some other alleles. How can we generalize using the results for a small number of HLA alleles? Of course, we can do an analysis for HLA-A2, but a systematic study requires the translation of results into poorly characterized alleles?

Margalit: I understand how the bias you mentioned affects the P1 position, because HLA-A2 prefers hydrophobic residues at the C-terminal position. But I don't see a problem with the flanking residue because there is no dependence between these two positions. It is detached; it doesn't exist any longer in the peptide after the cleavage, so I have a strong feeling that this is a proteasomal cleavage.

Brusic: Perhaps. In our unpublished studies we have seen clustering of potential targets in regions of the protein. T cell epitopes do cluster: hydrophobic regions are preferred by HLA-A2. This could introduce additional biases, compared with HLA-A3 binding peptides, which prefer charged residues at the C-terminus.

Kesmir: I agree with all of your comments that this is not really the right approach to tackle proteasome specificity with. On the other hand, if we use this approach and then test it on the real degradation data, we are able to predict 75% of the cleavages made by the immunoproteasome correctly. This is very good. For constitutive proteasome cleavages we can only predict 45%. We know about the bias problem and we are trying to reduce it as much as we can, but there is still some signal there that you can use to classify the cleavage sites from the non-cleavage sites. In five years, if we have sufficient degradation data that we can also understand the stochastic nature of these enzymes, then I think this will be fine.

Brusic: The key issue is to understand what can be done with data, and how can we improve our knowledge. This also touches on the issue about the need for reference data sets that are representative. If we got all data together and perform data preparation we can have richer data sets. The experimental work following on from predictions could be designed to also produce background information data. If we don't have many data, only simple predictive models can be built.

Gulukota: You mentioned that one of the possible false positives here is whether there is a cleavage site in the middle of the peptide. How many such cleavage sites would you find if you looked at the rest of the protein? You talked about aligning the peptide to the protein, and you are saying there is a cleavage site at both ends of the peptide, but there should not be one in the middle.

Margalit: There is a low cleavage score potentially in the middle.

Gulukota: How many sites would you find in the rest of the protein with low potential cleavage sites? The criterion appears pretty weak.

Margalit: It looks very weak, because we found only two positions at each side. We looked at 50 positions at each side and we didn't find any significance in any of the other positions. The degradation studies initially found additional significant positions at the N-terminal region of the cleavage site, and not in the C-terminal region. This fits nicely with the experimental results. There is a recent paper on the immunoproteasome in which Schild and colleagues also find that the main contribution is in the P1 and P1′ position, and with very weak contributions from other residues (Toes et al 2001). It is a very weak signal. I believe that other residues in the flanking regions have some influence, but we haven't found it yet. I wanted to raise for discussion the issue of how we integrate all this into a prediction scheme that integrates TAP binding and the MHC. We could do this, rank first the peptides by their cleavage potential and then by binding scores. The cleavage provides an additional parameter.

Beck: At the beginning you commented that the eluted peptide data can also be used to validate genome annotation or gene prediction. This is an interesting point, and it has not been taken up by the genome community so far. It probably also depends on how big the databases are that contain these sequences of experimentally determined peptides rather than predicted peptides.

Margalit: This was an exciting idea, and we wanted to do it, but we didn't end up with 100s of peptides. We had 500 peptides which we succeeded to match to human proteins and human genes. Most of them were known and had already been annotated. We were left with 30% of the 500 genes.

Beck: That's a significant proportion. If we could add even a fraction of this to the annotation, it would give a lot of additional value. How big is the data set of all known eluted peptides?

Stevanović: It is hard to be precise, but it is several hundred.

Beck: Presumably it will be difficult to get this figure into thousands of peptide sequences.

Stevanović: Too difficult.

Margalit: There are some interesting examples. We found two peptides that had a substitution in one position. We looked at the paper describing their discovery, and they were eluted from the same cell. We found that they matched two paralogue proteins. This means that these two paralogue proteins were both expressed in that cell.

Beck: Which is the best database to look for these eluted peptides?

Brusic: The best database of eluted peptides is SYFPEITHI.

Rammensee: You mentioned that the N-terminus of the peptides produced by the proteasome might be selected for fitting well to TAP. There is an arginine at the N-terminus which then binds well to TAP. But there are aminopeptidases also in the cytosol, and they would compete with TAP for the N-terminus of these peptides.

Margalit: This is all based on the premise that there is very little trimming in the cytosol.

Rammensee: We can't assume this. Jack Bennink and Jonathan Yewdell showed convincingly that a peptide brought into the cytosol is rapidly degraded. It can only be degraded by peptidases. We have to assume that the half-life of free peptides in the cytosol is very short. We cannot assume that they are running around for a long time.

Perelson: What do you mean by very short?

Rammensee: Two or three minutes, I suspect.

Margalit: So how would you explain this impressive compatibility between the preferences for the first residue with TAP and the P1′ position of the proteasome. It is an impressive coincidence.

Kesmir: One explanation is that people have shown that the immuno-proteasomes can co-localize with TAP.

Rammensee: In general, I would conclude from this that the specificity of TAP is the least important of the three components — proteasome, HLA and TAP.

Petrovsky: I can't see how you can say that TAP has the least specificity or is the least important. The proteasome is much less specific than TAP in terms of its cleavage sites and does not have a defined motif unlike TAP or MHC. Also, TAP is the major route into the MHC class I pathway and therefore acts as the gatekeeper. This is a highly important role as shown by the lack of class I expression when TAP is knocked out.

Rammensee: You need to distinguish between our ability to predict something and the mechanism. The proteasome is very conserved in its cleavage, the TAP also. Both have to work with very polymorphic MHC molecules. Maybe the

specificity of TAP is so well adapted to that of the proteasome that if the proteasome makes a peptide it is guaranteed that TAP collects it.

Petrovsky: That's a better way of putting it.

Margalit: So we don't need a predictive scheme for TAP.

Petrovsky: If TAP predicts proteosome and proteasome predicts TAP, and TAP prediction works much better, one could dispense with the proteasome prediction rather than the other way around.

Brusic: If proteasome cleavage is probabilistic, then this process will produce more of some peptides, but altogether it will produce all different peptides.

Kesmir: I didn't say it is probabilistic but it does have a stochastic component, so you can't reproduce your experiments 100%.

Brusic: When we examine all known HLA class I-binding peptides all amino acids will be represented at N-termini of the peptides.

Borras-Cuesta: If one assumes that the C-terminal amino acid is due to the cleavage by the proteasome, then in principal the C-terminal amino acid of all the different motifs should be hydrophobic. Is this the case?

Kesmir: No. You can't say that the proteasome only cleaves after hydrophobic residues. Consider that the proteasome was there to degrade the proteins. The best way to degrade a protein is to be as non-specific as possible. Then came the immunoproteasome, but there are some cells that express the immuno-proteasome all the time. The main function of the immunoproteasome is still to degrade the proteins. This is very important for the cell. It has preference for the hydrophobic residues but it still cleaves after other residues. If you make a kind of population study on your MHC binding markers, look at the diversity of amino acids on the ninth position and compare it with the second position, you will see that on the second position there are more diverse amino acids than on the ninth position.

Borras-Cuesta: If you have this type of processing, you would need mainly hydrophobic residues in the C-terminus.

Margalit: Why? It also cleaves after basic residues.

Borras-Cuesta: Wouldn't it be more logical to assume that it binds to MHC and is then trimmed off from the ends.

Rammensee: The proteasome is in a different place than the MHC. There are MHC molecules that like hydrophobic ends, those that like basic ends, those that like acidic ends (we are not aware of them because they are in the chicken).

Schönbach: One should also look at the amount of disorder in proteins to see whether there is any association between the preference for proteins cleaved by the proteasome and the degree of disorder. Disordered proteins seem to be more prone to degradation (Iakoucheva et al 2001). There is a preference for certain amino acids (e.g. PEST sequence).

Petrovsky: Isn't there an intrinsic contradiction here? Because you have done your analysis on MHC binding peptides, you are presuming that there is no trimming because otherwise your analysis is invalid. You are saying this is what is transported in, so pre-trimming must be occurring.

Margalit: There is no trimming after the proteasome cleavage.

Rammensee: Not at the C-terminus, but there is trimming at the N-terminus. That's what we agree on.

Brusic: What about all the reports that state that there are some peptides whose C-termini actually protrude out of the groove when they are bound to class I MHCs?

Rammensee: There was one peptide on B27 that was very long and that was recognized by antibody.

Brusic: There was another 10-mer peptide one that had glycine at position 10 protruding out of the groove.

Stevanović: In the experiments done in Hansjörg Schild's lab, several proteins and many precursor proteins have been digested by the proteasome. We really found that every amino acid that you can think of can be a cleavage site of the proteasome. We even found cleavage after and before proline and glycine. Hydrophobics are preferred, but cleavage can occur at any site.

Margalit: This is good, because the role of the proteosome is to degrade proteins.

Borras-Cuesta: What makes the difference? You end up with the hydrophobic residues. Whatever the mechanism, hydrophobic residues are found at the C-terminus of processed peptides.

Stevanović: The peptides bind to MHC molecules by hydrophobic residues. There is a very strong bias to hydrophobic C-termini. But the proteosome is creating peptides with any kind of C-terminal amino acid.

Petrovsky: This means that your predictions can never go beyond the bias. You will never go beyond 60% because you can't. Essentially you are saying that it could be cutting anywhere.

Rammensee: This is a matter of the amount of quantified data we put in. Then we get to a certain probability which will never be 100%.

Petrovsky: What it is saying is that it can cleave anywhere, so there is a slight bias which can be predicted, which gets us to 60%. But it doesn't matter how many additional data there are, they will show that the proteasome can ultimately cut anywhere although it is more likely to cut in some places than others.

Stevanović: If you include the quantitative data then it is just a question of distinguishing between specificity and sensitivity. If you set the threshold rather low then you get any cleavage perfectly, but you have 300 cleavages in a proteasome. If you minimize the potential cleavages in the proteasome by your prediction threshold to 10, then perhaps you miss 80% of all cleavages.

References

Iakoucheva LM, Kimzey AL, Masselon CD et al 2001 Identification of intrinsic order and disorder in the DNA repair protein XPA. Protein Sci 10:560–571

Toes RE, Nussbaum AK, Degermann S et al 2001 Discrete cleavage motifs of constitutive and immunoproteasomes revealed by quantitative analysis of cleavage products. J Exp Med 194:1–12

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

General discussion I

Rammensee: I'd like to welcome discussion of general points regarding the papers we have heard so far. What we can agree on is that we have a lot of data to work on, but the connectivity between the data and databases needs to be optimized. But we still don't have enough data, and sometimes it is hard to generate hard data to put into prediction models.

Kesmir: We started the day with a classification of immunoinformatics into two branches: soft and semi-soft. How can we join these two different approaches? Is it possible to combine the mathematical models and computer simulations of the immune system with the sequence analysis-based work? We should discuss this. Aminopeptidase activity that we have just been hearing about is a good example, because these peptidases don't have a great specificity, but it is a matter of how much they can access the peptide. This gives the end result of the trimming. We can only work out how long the peptides come into contact with peptidases if we make mathematical models of how peptides are generated, how they are transported to the endoplasmic reticulum (ER), how fast they bind to MHC and so on.

Rammensee: I think the answer to your question is rather easy. The test of any mathematical model is the experiment.

Kesmir: What I am saying is that with a mathematical model we can estimate how much peptide can be exposed to aminopeptidase, and then we can include this into our optimal epitope predictions.

Rammensee: This would be difficult.

Bernaschi: Hans-Georg Rammensee, from what you say it sounds like it is simple to test model predictions by experiments. My understanding is that it is very difficult to make a good experiment in immunology.

Rammensee: I wasn't suggesting that the experiments are easy.

Bernaschi: If I want to know the half-life of a T cell, for example, I don't have an easy way to answer this question. There should be more interaction between people in the lab and those making mathematical models. It is not as simple as one person making a model and then this being tested by an experimenter. Often it is not possible to make a good model because of a lack of specific data. Other times there are too many data, and it is difficult to identify the right ones to use. Slightly different approaches should be used to provide a methodology for

finding the right information required by people working on new models, and then the model should be tested.

Rammensee: As I tried to say earlier, for some modelling there is no way to prove its accuracy. For instance, modelling the half-lives of T cells. This belongs to the 'soft' branch of immunoinformatics.

Kesmir: Nonetheless, we can still use that estimate. It is better than just saying we don't know.

Brusic: There are deterministic and statistical questions that a researcher may ask. Some questions can be answered by specific experimental methods while others need to be treated statistically. We know very well that before elections pollsters take a sample of say 1000 voters and can predict the outcome of the election fairly accurately. In biology we can do the same for certain problems, but we need to know the limit of these predictions and how they can be applied. If we can do direct experimental validation, then it is a deterministic problem. The explanation of many deterministic measurements typically requires a statistical approach.

Bernaschi: This raises another interesting point. What is the meaning of statistics? If you consider that you are working in a field with 40 million people with a specific disease and you build your statistical models from just 10 people, does this make any sense? The answer is no. But it is the only thing you can do in an experiment.

Rammensee: It depends on the difference between the two groups, I guess, and the type of experiment.

Lefranc: Nick Petrovsky, I was quite interested by your slide describing all the steps needed for sharing experimental and clinical data between labs and clinicians. At the end of your slide there were two headings. One was managing laboratory information and the other was protecting clinical data. These are two areas where a lot of work remains to be done. Could you comment on the kinds of standards to be set up?

Petrovsky: What I was saying was that it is very hard to assess information coming out of a lot of different laboratories. Although in their publications people are meant to describe enough information in their methods for other people to be able to reproduce those experiments, we all know this can be very difficult. This is mainly because if people said exactly what they did, the methods section would end up being huge. Similarly, with clinical data the problem is that they are generally very incomplete. If you go to any clinical databases they are very much biased by the clinician and their ideas about the disease. In fact, if you talk to two clinicians they may actually define the disease quite differently. As we get laboratory data, such as those provided by gene expression arrays, it will be hard to match this precise information up with clinical information if the latter is imprecise. We are saying that we need to ensure that both sets of information are precise and standardized. At that point it will be possible to bring the two together.

People are currently trying to bring together expression data and disease data, and one of the reasons they are finding this problematic is that there is so much imprecision and lack of standardization in both data sets. We need to try to introduce standards and get some consistency in diagnostic criteria and clinical attributes so that we can interpret the laboratory data in a more consistent way.

DeLisi: Expression arrays can achieve this: they can stratify diseases that were previously thought to be the same disease. Particularly with neurological diseases this is a serious problem. Imaging will help there, but one needs to find hard phenotypic correlates of what is going on genetically.

Lefranc: To make the clinical data more precise, any available phenotypic, serological or genetic markers related to a gene should be entered. At the beginning of the 1980s when we sequenced the immunoglobulin *IGHG* and *IGHA* genes, we cloned and sequenced genes from individuals for whom we had previously analysed the familial pedigree and determined the Gm allotypes by serological typing. In many cases, unfortunately, a lot of information was lost because many labs which cloned and sequenced genes at that time were not concerned by the genetic information and polymorphisms (serological, RFLP, etc.) associated with their clone or phage sequences. Coming back to the clinical side, what kind of standard information do you see for the future? What is the minimum level of information that needs to be collected?

Petrovsky: If you are researching a disease from the laboratory viewpoint, you need to ensure the clinicians you are working with are able to give you the classification of disease that they are using. Increasingly, clinicians are trying to agree on common diagnostic criteria. Someone who runs an assay has to be able to reference it in order to publish it. If a clinician tells you that this is a group of patients with a particular type of rheumatoid arthritis, you should demand a reference for how they were classified as being in that particular subgroup. If they can't, then you have a problem. Many clinical groups have decided that they need a system for classifying particular diseases and have developed internal guidelines on classification. Once a group of experts has agreed on a system, then everyone else generally eventually adopts it.

Kellam: This already exists in some diseases, for example the lymphomas and leukaemias, which have international recognized standards for diagnosis and classification.

Petrovsky: It is like annotation. As long as they can say which guidelines they are using, people can then go back to the source and work out what they are dealing with. As long as they can reference the source, that should be fine. As scientists we will have very noisy data if the person giving us the clinical samples we are using in our studies isn't classifying them according to some sort of defined criteria.

Gulukota: In some of the microarray communities they use strict classifications such as that given in the commercial package SNOMED which has an ontological

classification for much of medicine i.e. pathologies, tissue anatomy, drugs etc. Something of this sort could be developed in the open standards community in a manner similar to Gene Ontology.

Kellam: I think SNOMED is designed to be this: the equivalent clinical description to Gene Ontology.

Gulukota: The problem with SNOMED is that it isn't free.

Kellam: There is an open-source equivalent available. There are some microarray pages as well for clinical annotation. Again, it is difficult to get everything annotated retrospectively.

De Groot: The problem is that most of the clinical information is hand-written in doctors' scribble.

Petrovsky: SNOMED is more a dictionary than it is a set of diagnostic criteria for each disease. It still leaves the diagnosis to the clinician in their individual judgement, which is not annotatable, unless you annotate the name of the diagnosing clinician!

Gulukota: I agree. Often clinical trials don't just say what the disease is but also have explicit inclusion and exclusion criteria. These are fairly rigorous. We might need to have something like this in mind when we are investigating a particular disease from a collaboration point of view.

Rammensee: Our task is to talk to the clinicians and tell them what kind of information we need for our different purposes. I don't think we can generalize about the conditions which our clinical partners have to follow. We are not the right people to do this.

De Groot: What will happen is that as they come to us for an explanation, we will say they need to start collecting HLA data if you want us to explain why your therapeutic proteins are causing side effects, for example. There will be an evolution and this will be important. I think also that there is an acceptance of immunoinformatics which is key here. Vladimir Brusic pointed out that a few years ago we were looking at a black box. There has been a change in the acceptance of this technology so we are now in a position to start asking for some better data.

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

Computational vaccinology: quantitative approaches

Darren R. Flower, Helen McSparron, Martin J Blythe, Christianna Zygouri, Debra Taylor, Pingping Guan, Shouzhan Wan, Peter V. Coveney*[1], Valerie Walshe, Persephone Borrow and Irini A. Doytchinova

*Edward Jenner Institute for Vaccine Research, High Street, Compton, Berkshire, RG0 7NN and *Centre for Computational Science, Department of Chemistry, Queen Mary, University of London, Mile End Road, London E1 4NS, UK*

Abstract. The immune system is hierarchical and has many levels, exhibiting much emergent behaviour. However, at its heart are molecular recognition events that are indistinguishable from other types of biomacromolecular interaction. These can be addressed well by quantitative experimental and theoretical biophysical techniques, and particularly by methods from drug design. We review here our approach to computational immunovaccinology. In particular, we describe the JenPep database and two new techniques for T cell epitope prediction. One is based on quantitative structure–activity relationships (a 3D-QSAR method based on CoMSIA and another 2D method based on the Free–Wilson approach) and the other on atomistic molecular dynamic simulations using high performance computing. JenPep (*http://www.jenner.ac.uk/JenPep*) is a relational database system supporting quantitative data on peptide binding to major histocompatibility complexes, TAP transporters, TCR-pMHC complexes, and an annotated list of B cell and T cell epitopes. Our 2D-QSAR method factors the contribution to peptide binding from individual amino acids as well as 1–2 and 1–3 residue interactions. In the 3D-QSAR approach, the influence of five physicochemical properties (volume, electrostatic potential, hydrophobicity, hydrogen-bond donor and acceptor abilities) on peptide affinity were considered. Both methods are exemplified through their application to the well-studied problem of peptide binding to the human class I MHC molecule HLA-A*0201.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 102–125

One of the principal goals of immunoinformatics, a newly emergent branch of bioinformatics focusing on immunobiology problems, is to develop computational vaccinology (computer-aided vaccine design or CAVD) as a practical science for the discovery of new vaccines. The recognition of antigenic

[1]Present address: Christopher Ingold Labs, Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ.

epitopes, either small discrete T cell epitopes or large conformational epitopes recognized by soluble antibodies or B cells, is the key molecular event at the heart of the immune response. Within the context of cellular immunology, peptide immunogenicity is contingent upon the ability of epitopes to bind major histocompatibility complexes (MHC) and to be recognized subsequently by T cell receptors (TCR). Traditionally, T cell epitopes have been identified by examining immune responses to overlapping peptides generated from target antigens. Logistically, this process becomes prohibitive when studying the thousands of gene products found within microbial genomes, and recourse to computational analysis is required to reduce subsequent experimental work. It is well known that only peptides making high affinity interactions with MHC molecules are recognized as T cell epitopes (Sette et al 1994). In terms of a competition assay, the IC_{50} must be less than 500 nM. Thus MHC-binding prediction is a necessary preliminary to the identification of T cell epitopes. The accurate prediction of B and T cell epitopes, around which modern polyepitope vaccines are constructed, is a pivotal challenge for CAVD. While the prediction of B cell epitopes remains primitive (Alix 1999), or depends on an often-elusive knowledge of protein structure (Thornton et al 1986), a broad spectrum of sophisticated methods for the prediction of T cell epitopes has evolved (Flower et al 2002). These began with early motif methods (Sette et al 1989), and have grown to exploit both qualitative or semi-quantitative approaches, typified by neural network classification methods (Honeyman et al 1998), and a variety of more quantitative approaches (Parker et al 1994, Rognan et al 1999, Doytchinova & Flower 2002a).

We review here our quantitative approach to the rapidly evolving field of computational vaccinology, and include discussion of recent updates to our JenPep database and the application of two powerful techniques for T cell epitope prediction. One is based on a quantitative structure–activity relationship, or QSAR, approach, implementing both 2D and 3D methods (Doytchinova & Flower 2001, 2002b, Doytchinova et al 2002), and the other on atomistic molecular dynamics simulations using high performance computing.

JenPep

JenPep is an integrated relational database system for functional and quantitative data on peptide binding within immunobiology (Blythe et al 2002) and is the first of its type to concentrate on thermodynamic measurements, thus complementing the existing system. The database is available free via the Internet (*http://www.jenner.ac.uk/JenPep*). The current version of JenPep (version 2.0) is composed of five types of data: (i) quantitative measures for peptides binding to class I and class II MHC; (ii) a compendium of T-cell epitopes; (iii) quantitative

TABLE 1 Epitope data

Peptide class	*Total number of peptides*	*Length*[a] *distribution*	*Class I*[b]	*Length*[a] *distribution*	*Class II*[b]	*Length*[a] *distribution*
B cell Epitope	816	3–47				
T cell Epitope	3218	7–35	2060	7–24	1158	8–35
MHC binding	12336	4–28	6411	4–23	5925	7–28
TCR–pMHC	49	8–20				
TAP Transporter	441	7–15				

[a] Range in amino acids
[b] Number of peptides

measures for peptide binding to the TAP peptide transporter; (iv) affinity measures for the formation of the peptide–MHC–TCR complex; and (v) a compilation of B cell epitopes. The size of the database is outlined in Table 1. Version 2.0 is implemented in a bespoke system using open source postgreSQL as the database engine and a graphical user interface (GUI) written in perl/HTML. Together with the peptide sequence, JenPep includes various kinds of binding measure, MHC restriction and the protein sequence from which the peptide originates. Data on T and B cell epitopes are currently limited to a list of binders and, in this context, we rely on the judgement of experimental immunologists to define what are, or are not, epitopes.

We should like to extend JenPep to allow analysis of non-natural mutants of MHC molecules and non-amino acid ligands of MHC molecules, such as post-translationally modified peptides, as well as complementing our thermodynamic data with kinetic data on peptide binding. Another addition to our cellular immunological data would be information on other immunological recognition events, such superantigen binding to MHCs and TCRs and the interaction of cell surface co-receptors.

We also look forward to the day when immunologists submit their experimental binding data to an online archive, such as ours, much as today's molecular biologists must submit their data to a publicly curated sequence database. Taking a lead from the Interpro Project (Apweiler et al 2002), one can envisage an international collaboration aimed at producing a broadly focused immunogenicity database. In Interpro, existing databases of sequence families, such as PRINTS (Attwood et al 2002), have been combined to produce a more complete coverage of known sequence families, combining annotation details from the different component databases. A similar super-database, incorporating, *inter alia*, JenPep, SYFPEITHI (Rammensee et al 1999), the HIV Molecular Immunology database

(Korber et al 2002), and FIMM (Schönbach et al 2000), into a comprehensive database of immunogenic peptides, is the obvious immunological counterpart to Interpro.

QSAR methods for binding affinity prediction

QSAR methods are powerful tools for the prediction and rationalization of structure-property relationships within physical science, and have proved particularly successful within pharmaceutical research. The fundamental objective of QSAR is to take a set of molecules, for which a biological response has been measured, and using statistical, or artificial intelligence methods, such as an artificial neural network or genetic algorithm, relate this measured activity to some description of their structure. The outcome, then, of a QSAR study are equations that relate, through statistically sound and hopefully predictive models, the activity, or, more generally, the biological responses or physical properties, of a set of molecules to their molecular structure. Their ability to provide mechanistic explanations is dependent on the form of the particular molecular description. There are two areas of technical development with QSAR: the development of new, and hopefully improved, descriptions of molecular structure and the development of new statistical or artificial intelligence methods which can relate these descriptions to some measured biological or physical property. We have developed or applied two techniques from QSAR. One is based solely on the sequence of peptides, this is a 2D QSAR technique which we call the Additive method (Doytchinova et al 2002). The second technique is based on CoMSIA (Klebe et al 1994), and is a 3D QSAR technique using the 3D coordinates of bound peptides (Doytchinova & Flower 2001, 2002b).

The additive method exploits the Free–Wilson concept, a well-established QSAR technique (Free & Wilson 1964), whereby each substituent makes an additive and constant contribution to the biological activity irrespective of structural variation in the rest of the molecule. The independent binding of sidechains (IBS) hypothesis, developed by Parker et al (1994) is the immunological counterpart to the Free–Wilson concept. We have extended this concept by adding additional terms that account for near neighbour side-chain interactions. The binding affinity of a peptide will depend on contributions from each amino acid as well as the interactions of adjacent and every second side-chain:

$$\textit{binding affinity} = \textit{const} + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_i P_{i+1} + \sum_{i=1}^{7} P_i P_{i+2}, \qquad (1)$$

where the *const* accounts, albeit nominally, for the peptide backbone contribution, $\Sigma_{i=1}^{9} P_i$ is the sum of amino acids contributions at each position, $\Sigma_{i=1}^{8} P_i P_{i+1}$ is the sum of adjacent peptide side-chain interactions, and $\Sigma_{i=1}^{7} P_i P_{i+2}$ is the sum of every second side-chain interaction. This choice of parameterization is based on the observation that HLA-A*0201-bound peptides assume extended but twisted conformations, so that adjacent side-chains point in essentially opposite directions: both 1–2 and 1–3 interactions are possible between side-chains.

In our initial application, we extracted from JenPep 420 IC_{50} values for 340 nonamer peptides sequences that bound to the HLA-A*0201 molecule. As is common practice amongst QSAR practitioners, IC_{50} values were converted to p-units (negative decimal logarithm). The mechanistic details of the additive method are outlined in Fig. 1. A term is equal to 1 when a certain amino acid at a certain position, or a certain interaction, exists, and 0 otherwise. As the columns are more numerous than the rows, the equations were solved using PLS and their predictive power assessed using cross-validation and multiple linear regression parameters (see Table 2). The contributions made by individual amino acids and by certain interacting side-chains at particular peptide positions are shown in Fig. 2. We have subsequently applied this method to a variety of other alleles, the results of which will be published separately. The statistics for these models are shown in Table 3. Moreover, we have developed an internet server, called MHCPred, which implements the additive method. It is available over the Internet at *http://www.jenner.ac.uk/MHCPred* (see Fig. 3).

One of the most reliable methods for investigating the structure–activity trends within sets of biological molecules is 3D-QSAR. The explanatory power of 3D-QSAR methods is considerable, manifest not only in their accurate prediction of binding affinities, but through their capacity to display advantageous and disadvantageous 3D interaction potential mapped onto the molecular structure being investigated. We have applied a 3D-QSAR method (CoMSIA) to gain an understanding of the relationship between certain physicochemical properties (volume, electrostatic potential, local hydrophobicity and hydrogen-bond donor and acceptor abilities) and the affinities of HLA-A*0201 binding peptides.

266 9-mer peptides were analysed using CoMSIA. As before, their IC_{50} values were taken from JenPep and converted to p-units. All molecular modelling and QSAR calculations were performed on a Silicon Graphics octane workstation using the SYBYL 6.7 molecular modelling software. The X-ray structure of the nonameric viral peptide TLTSCNTSV[39] was used as the template onto which all structures were built. Five similarity indices were calculated, using a common probe atom with 1 Å radius, charge +1, hydrophobicity +1, hydrogen-bond donor and acceptor properties +1. The predictive power of the final model was assessed using the same statistical parameters as for the additive method. Leave-one-out

Peptides	pIC_{50}	20AA×9 180 columns	20×20×8 3200 columns	20×20×7 2800 columns
1	4,602	1.....0.....1	1.....0.....1	1.....0.....1
.....	.....	...1.....1....	...1.....1....	...1.....1....
420	8,959	0.....1.....0	0.....1.....0	0.....1.....0

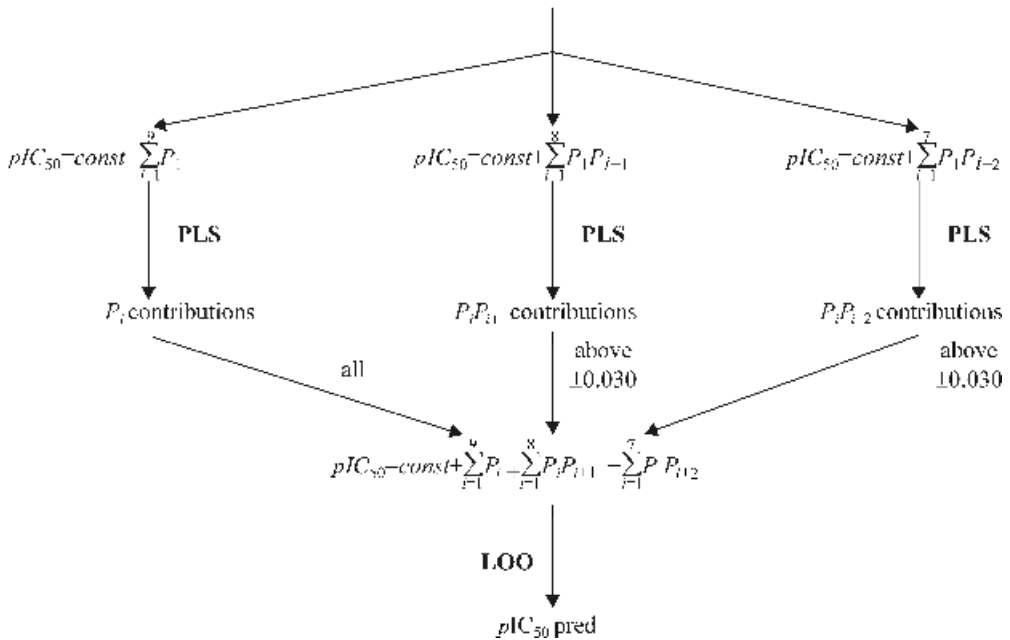


FIG. 1. Additive method algorithm.

TABLE 2 Statistical parameters for the HLA-A*0201 additive model

	Number of peptides = 340	
	Number of components = 5	
	$q^2 = 0.337$	
	SEP = 0.726	
	$r^2 = 0.898$	
	$SEE = 0.285$	
	$F = 588.883$	
RESIDUALS:		
< 0.5		172 peptides 50.5 %
0.5–1.0		128 peptides 37.5 %
> 1.0		40 peptides 12.0%
	|Residual| = 0.573 (SD = 0.442)	

cross-validation (LOO-CV), CV in two and five groups, and a bootstrap analysis (20 runs) were performed.

The model was improved by excluding a limited number of very poorly predicted peptides in a stepwise manner, beginning with the peptide with the highest residual. The final CV model had significantly higher parameter values: q^2 = 0.683 at 7 components and r^2 = 0.891. This model was used to predict the binding affinities of the excluded peptides. The predictions improved for both well-predicted and poorly predicted peptides. The mean |*residual*| value for this model was 0.489. The stability of the final CoMSIA model was tested by CV in two and five groups. The mean q^2 for 20 runs for CV in five groups was 0.656, which is close to the LOO-CV value. The 'leave-half-out' CV gave a lower value for q^2 (the mean of 50 runs is 0.558), which remains close to the other q^2 values, and $r^2_{\text{bootstrap}} = 0.924$. The non-cross-validated CoMSIA model was used to display the coefficient contour maps. Results were visualized using the 'StDev*Coeff' mapping option contoured by actual values (see Fig. 4). We have subsequently applied this approach to other alleles, the results of which we will publish separately. The statistics for our CoMSIA models are shown in Table 4.

Atomistic molecular dynamic simulations

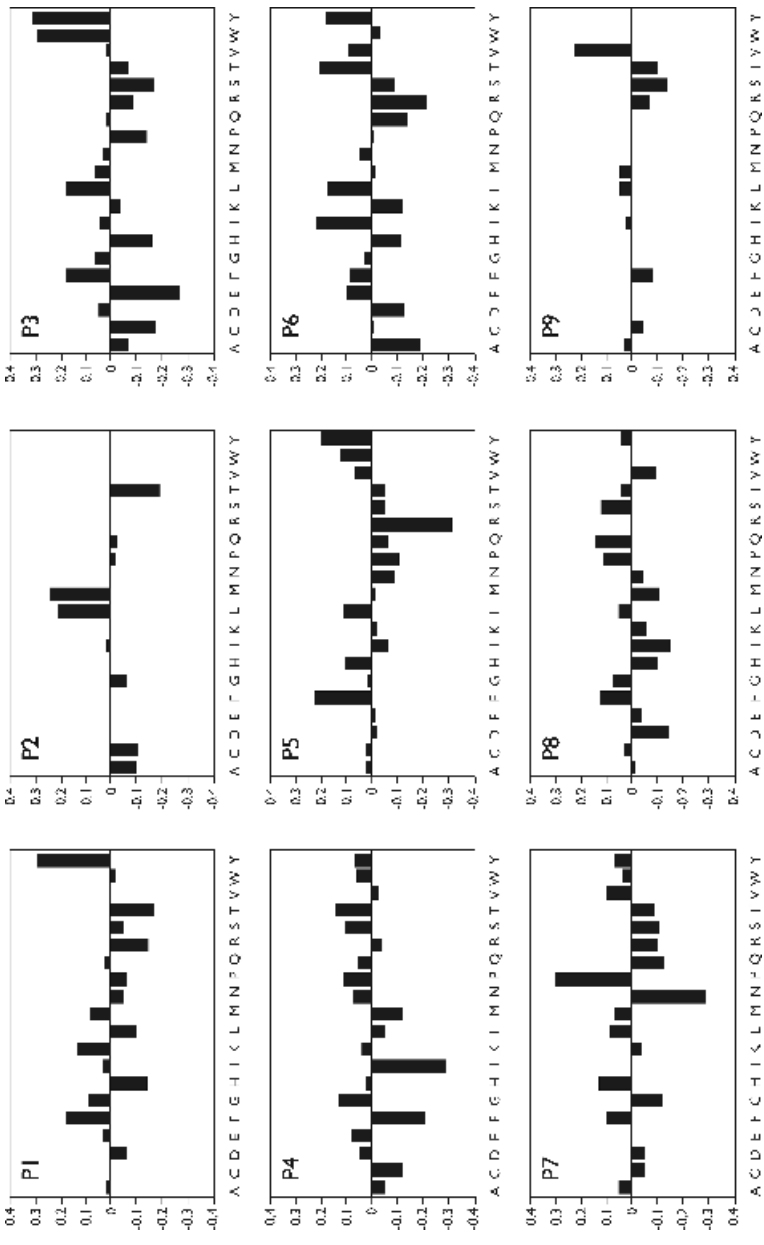
The growth of computer power during the last two decades has allowed biologically interesting systems to be studied using atomistic molecular dynamics methodology. We are still faced, however, with problems concerning the short duration of simulations possible on current serial machines. Many approaches have been tried to circumvent these limitations, but only with restricted success.

Almost any attempt to reach longer time scales will result in more approximations in the model. Previous attempts to utilize molecular dynamics and other atomistic simulation methods to investigate peptide MHC interactions have foundered on such technical limitations. Many methods exist which predict thermodynamic properties from simulations, but typically take an unrealistically long time: simulations yielding a free energy of binding require at least 10 nanoseconds. An average desktop serial workstation requires a compute time in the order of 300 hours per nanosecond. With down-time, simulating a few dozen peptides might occupy a machine for several years.

To escape these limitations, we might take advantage of high performance, massively parallel implementations of molecular dynamic (MD) codes running on supercomputers with 128, 256 or 512 nodes. We are pursuing this goal within the context of grid computing: an ambitious global effort to develop an environment where individual users access computational or data resources simply and transparently, irrespective of their location, and which is named by analogy with the national power transmission grid. If one desires to switch on a light or run a refrigerator, one is not required to wait while sufficient current is downloaded, thus grid computing seeks to make available all necessary compute power at the point of need. As part of the RealityGrid Project (*http://www.realitygrid.org*), we are using an implementation of the popular molecular dynamics force-field AMBER, as implemented in LAMMPS, a specially written parallel molecular dynamics program, to simulate solvated A*0201 peptide complexes as an initial test of this approach. Biomolecular simulations show significant acceleration relative to single processor runs (see Table 5), reducing the time needed to simulate a nanosecond to 12 hours. We intend to utilize such performance gains to run large simulations for a sufficient duration that atomistic simulation of peptide–MHC complexes will become a realistic tool in epitope prediction.

Discussion

We have described the continuing development of quantitative approaches to the prediction of MHC binding built around data in JenPep, our database of binding measures. Despite the differences between the additive method and CoMSIA, we found good agreement between results generated by these techniques. CoMSIA can extrapolate, predicting the affinity of a peptide with an amino acid absent in the initial training set, but it can not assess the relative contribution of individual amino acids nor the interactions between them. The opposite is true for the additive method: it extrapolates poorly, but gives a good assessment of the contributions made by amino acids. Thus the results of our methods are complementary: they give greater insight when used together.


P1
P2
P3
P4
P5
P6
P7
P8
P9
A C D E F G H I K L M N P Q R S T V W Y

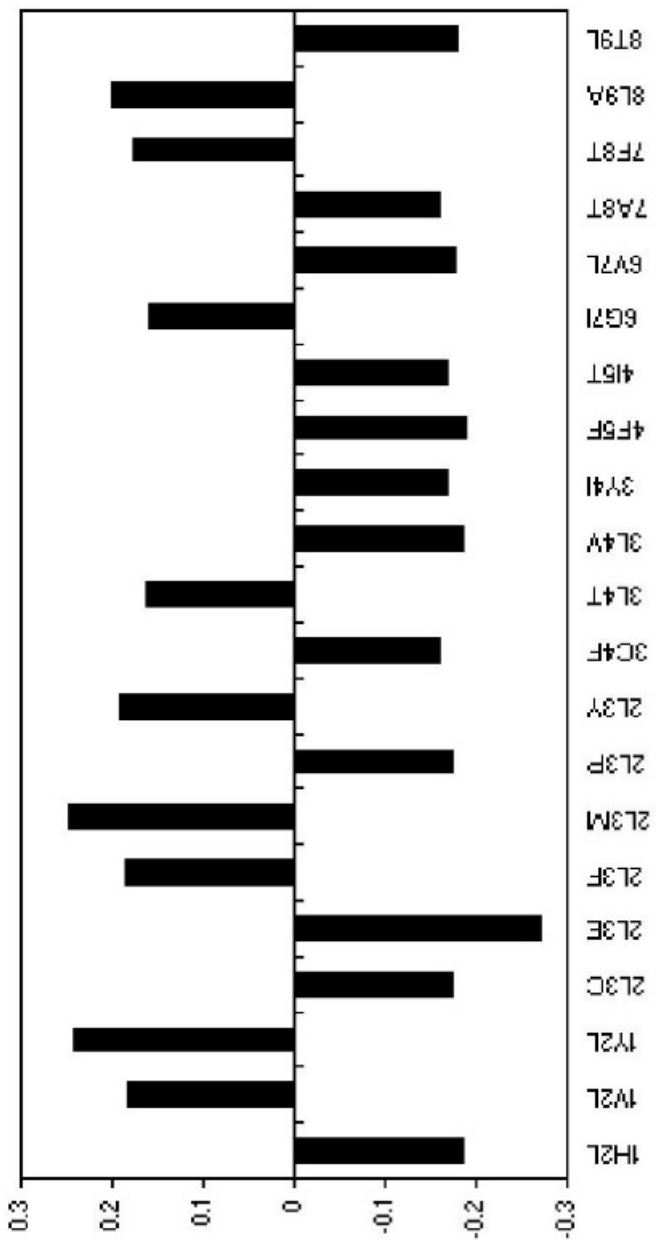


FIG. 2. Amino acid contributions for HLA-A*0201. (A) Amino Acid contributions. Contributions for all 20 amino acids at the nine positions of a 9-mer peptide binding to HLA-A*0201. Positions marked P1–P9. (B) Contributions of 1–2 interactions. A set of positive or negative high scoring pairwise side chain–side chain interactions between adjacent residues. Interactions are coded: 1G2H is a glycine at position 1 interacting with histidine at position 2. (C) 1–3 interactions. A set of positive or negative high scoring pairwise side chain-side chain interactions between residues one sequence position apart. Interactions are coded: 3S5W is a serine at position 3 interacting with tryptophan at position 5.

TABLE 3 Model statistics for additive method

	*A*0101*	*A*0201*	*A*0202*	*A*0203*	*A*0206*
n	95	335	69	62	57
q^2	0.42	0.377	0.317	0.327	0.475
NC	4	6	9	6	6
SEP	0.907	0.694	0.606	0.841	0.576
r^2	0.997	0.731	0.943	0.963	0.989
	*A*0301*	*A*1101*	*A*3101*	*A*6801*	*A*6802*
n	70	62	31	37	46
q^2	0.305	0.428	0.453	0.370	0.500
NC	4	3	6	4	7
SEP	0.699	0.593	0.727	0.664	0.647
R^2	0.972	0.977	0.990	0.974	0.983
	*B*3501*				
n	50				
q^2	0.516				
NC	8				
SEP	0.725				
R^2	0.996				

In developing these methods, we have encountered technical challenges that are only rarely encountered in QSAR analyses of small molecules. These include the number of molecules investigated, perhaps 10 times that for a small molecule study; the very size of the peptide molecules being studied; and the great diversity of physicochemical properties associated with each peptide. Since it is clear from crystallography that there are only minor differences in backbone conformation for nonamer peptides, we have avoided the thorny issue of molecular alignment in our CoMSIA studies by assuming a constant backbone. As most peptides are well predicted, variations in the binding conformation do not seem significant.

We have also extended these methods to address a number of related problems in immunobiology. For example, we have used both the CoMSIA and additive methods to refine A2 and A3 peptide-binding super-types (Doytchinova & Flower 2002c, 2003, Guan et al 2003a,b). We are also using in-house experimental cell surface stabilization assays to test out the predictivity of our modelling approach (Walshe, Doytchinova, Borrow and Flower, unpublished),

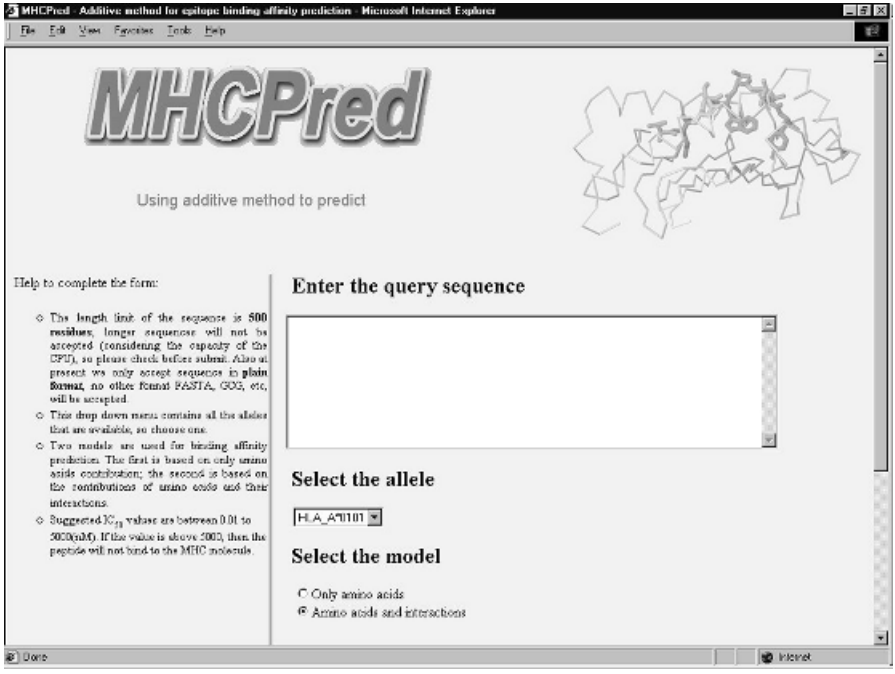


FIG. 3. MHCPred homepage.

and to this end we are designing and testing synthetic super-binding peptides as well as developing comprehensive models for poorly characterized alleles using experimental design. Furthermore, we are also applying our techniques to the iterative optimization of heteroclitic peptides as potential cancer vaccines (Rigley and Flower, unpublished).

Conclusions

Hitherto, a dichotomy of approaches has been apparent in immunoinformatics. Initially, work in the area was informed by an engineering or computer science perspective emphasizing the need to solve problems or reach objectives. This is reflected in the qualitative, or more precisely, semi-quantitative approach taken by the creators of databases such as MHCPEP (Brusic et al 1998) or SYFPEITHI (Rammensee et al 1999) or the users of neural networks as a prediction engine. A classification scheme is used here as a data fusion device to agglomerate and simplify accumulated epitope data. Recently, there has been a move, within the discipline, towards a quantitative approach (Rognan et al 1999, Doytchinova & Flower 2002a). To some extent this is grounded in a more physicochemical

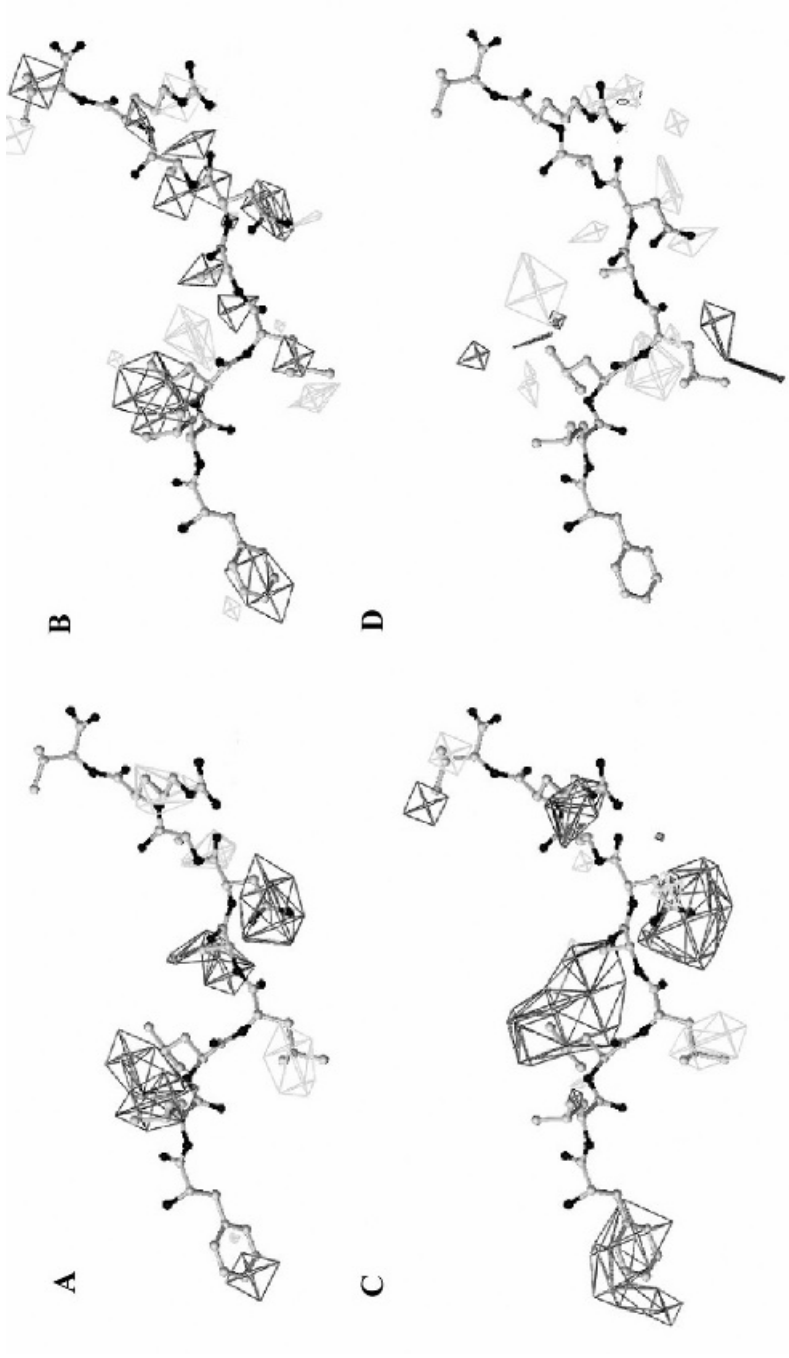

B
D
A
C

FIG. 4. CoMSIA stdev*coeff contour plots. (A) Steric map. (B) Electrostatic map. (C) Hydrophobic map. (D) H-bond donor ability. The peptide FLLLADARV ($pIC_{50}=8.620$) is shown. Black and grey polyhedra indicate regions where steric bulk, electron density, hydrophobicity and H-bond donors will enhance or decrease the affinity, respectively. Volume is well tolerated at P1, as well as at P2, P3, P5 and P6, while disfavoured areas exist at P4, P7 and P8. Negative potential is generally favoured, except between P3 and P5, and at P8. Areas of favourable local hydrophobicity exist at P1, P3, P5, P6, P8 and P9. Favoured hydrophilic groups are at P4 and P9. Hydrogen bond donors on the ligand are favoured at P3 and at P4 and acceptors on the ligand are favoured at P4, P6 and P8. The map of H-bond acceptor ability is not shown. Favoured H-bond acceptor areas exist at P4 and P8, disfavoured at P2, P3, P5, P6 and P7.

TABLE 4 Statistics for CoMSIA models

Parameter	*A*0201*	*A*0202*	*A*0203*	*A*0206*	*A*6802*
n	236	63	60	54	45
q^2_{loo}	0.683	0.534	0.621	0.523	0.385
NC	7	8	6	12	4
SEP	0.443	0.509	0.595	0.505	0.652
r^2	0.891	0.935	0.966	0.991	0.944
SEE	0.260	0.190	0.179	0.071	0.197
F *ratio*	265.082	97.199	247.303	363.764	168.149
Fractions:					
• Steric	0.145	0.144	0.162	0.110	0.114
• Electrostatic	0.320	0.228	0.177	0.282	0.265
• Hydrophobic	0.210	0.291	0.270	0.304	0.260
• H-bond donor	0.161	0.210	0.235	0.210	0.214
• H-bond acceptor	0.164	0.127	0.157	0.094	0.148
Residuals:					
• $res. \leqslant \lvert 0.5 \rvert$	166 70%	42 67%	44 73%	36 63%	27 60%
• $\lvert 0.5 \rvert < res. \leqslant \lvert 1.0 \rvert$	70 30%	21 33%	9 15%	19 33%	16 36%
• $res. > \lvert 1.0 \rvert$	0 0%	0 0%	7 12%	2 4%	2 4%
• mean |residual|	0.340	0.393	0.434	0.443	0.519
• standard deviation	0.246	0.281	0.356	0.310	0.332

		*A*1101*		*A*0301*		*A*3101*		*A*6801*	
Number of peptides		59		69		30		39	
q^2_{loo}		0.496		0.486		0.700		0.570	
Number of components		8		6		4		10	
SEP		0.588		0.629		0.551		0.655	
r^2		0.972		0.959		0.921		0.990	
	SEE	0.141		0.177		0.282		0.100	
F ratio		167.666		241.818		73.177		247.715	
	Steric	0.114		0.104		0.071		0.126	
	Electrostatic	0.234		0.277		0.254		0.250	
Fraction	Hydrophobic	0.250		0.260		0.225		0.280	
	H donor	0.261		0.226		0.364		0.237	
	H acceptor	0.141		0.133		0.087		0.107	
res. |0.5|		40	67.80%	42	60.87%	27	90%	23	58.97%
|0.5| < *res.* |1.0|		15	25.42%	20	27.03%	3	10%	14	35.90%
res. > |1.0|		4	6.78%	7	9.46%	0	0	2	5.13%
mean |residual|		0.443		0.585		0.179		0.440	
standard deviation		0.343		0.500		0.188		0.343	

TABLE 5 Biomolecular simulations show significant acceleration relative to single processor runs

Processor Number	*Speed up*
1	1.00
2	2.07
4	4.42
8	9.19
16	17.79
32	30.58
64	52.02
128	83.23

perception, with a focus on the use of direct data, such as IC_{50}s, and a greater implicit emphasis on explanation, and ultimately, a greater explicit understanding of underlying molecular mechanisms. Both viewpoints are increasingly immunologically aware, and are best seen as complementary. To some extent, the remaining conflicts between these differing perspectives can, in principal, be reconciled by methods originating, or finding application, in Drug Design, such as QSAR or molecular dynamics. They meet both objectives: seeking to explain and understand without sacrificing the ultimate utilitarian value of the undertaking.

The immune system is hierarchical and many levelled, exhibiting much emergent behaviour. However at the heart of the phenomena are straightforward molecular recognition events that are indistinguishable from other types of biomacromolecular interaction, such as enzyme–inhibitor or antagonist–receptor interactions. Binding of an epitope to a MHC or pMHC to a TCR is, at the level of underlying physicochemical phenomena, identical, say, to the binding of a drug to a receptor protein. Indeed, the terms agonist and antagonist, commonly used within the immunological community, originate from pharmacology.

An important corollary to this observation is the emphasis placed on the important role of non-anchor residues in influencing the energetics of peptide–MHC binding. In contradiction to the dogma extant amongst many immunologists, it is clear that anchor residues alone cannot account for peptide binding. Rather it is the combination, albeit weighted, of all amino acids within the peptide that ultimately determines the observed affinity of binding. Thus it is only methods which account for all interactions within a quantitative setting that can properly address the issue of binding prediction.

As part of its ambitious programme, the Edward Jenner Institute for Vaccine Research is committed to the development of computational vaccinology as a vital component in the battle against infectious disease. Informatics techniques have proved their worth time and again in the search for new drugs. The time is approaching when they will do the same for vaccine design. Methods that allow us to predict accurately individual epitopes or immunogenic proteins will prove to be crucial tools for the vaccinologist of tomorrow.

Acknowledgements

We should like to thank Prof Peter Beverley, Dr Vladimir Brusic, Dr Helen Kirkbride, Paul Taylor and Kelly Paine for useful discussions. We thank the EPSRC for funding the *RealityGrid* e-Science Pilot Project.

References

Alix AJ 1999 Predictive estimation of protein linear epitopes by using the program PEOPLE. Vaccine 18:311–314

Apweiler R, Attwood TK, Bairoch A et al 2001 The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29:37–40

Attwood TK, Blythe MJ, Flower DR et al 2002 PRINTS and PRINTS-S shed light on protein ancestry. Nucleic Acids Res 30:239–241

Blythe MJ, Doytchinova IA, Flower DR 2002 JenPep: a database of quantitative functional peptide data for immunology. Bioinformatics 18:434–439

Brusic V, Rudy G, Harrison LC 1998 MHCPEP, a database of MHC-binding peptides: update 1997. Nucleic Acids Res 26:368–371

Doytchinova IA, Flower DR 2001 Toward the quantitative prediction of T-cell epitopes: CoMFA and CoMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. J Med Chem 44:3572–3581

Doytchinova IA, Flower DR 2002a Quantitative approaches to computational vaccinology. Immunol Cell Biol 80:270–279

Doytchiniva IA, Flower DR 2002b Physicochemical explanation of peptide binding to HLA-A*0201 major histocompatibility complex. A three-dimensional quantitative structure–activity relationship study. Proteins 48:505–518

Doytchinova IA, Flower DR 2002c A comparative molecular similarity index analysis (CoMSIA) study identifies an HLA-A2 binding supermotif. J Comput Aided Mol Des 16:535–544

Doytchinova IA, Flower DR 2003 The HLA-A2 supermotif: a QSAR definition. Org Biomol Chem 1:2648–2654

Doytchinova IA, Blythe MJ, Flower DR 2002 An additive method for the prediction of protein–peptide binding affinity. Application to the MHC Class I molecule HLA-A*0201. J Proteome Res 1:263–272

Free SM Jr, Wilson JW 1964 A mathematical contribution to structure–activity studies. J Med Chem 7:395–399

Flower DR, Doytchinova IA, Paine K et al 2002 Computational vaccine design. In: Flower DR (ed) Drug design: cutting edge approaches, Royal Society of Chemistry, p 136–180

Guan P, Doytchinova IA, Flower DR 2003a HLA-A3 supermotif defined by quantitative structure-activity relationship analysis. Protein Eng 16:11–18

Guan P, Doytchinova IA, Flower DR 2003b A comparative molecular similarity indices (CoMSIA) study of peptide binding to the HLA-A3 superfamily. Bioorg Med Chem 11:2307–2311

Honeyman MC, Brusic V, Stone NL, Harrison LC 1998 Neural network-based prediction of candidate T-cell epitopes. Nat Biotechnol 16:966–969

Klebe G, Abraham U, Mietzner T 1994 Molecular similarity indices in a comparative analysis CoMSIA of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146

Korber BTM, Brander C, Haynes BF et al (eds) 2001 HIV molecular immunology. Los Alamos National Laboratory: Theoretical Biology and Biophysics, Los Alamos, New Mexico

Parker KC, Bednarek MA, Coligan JE 1994 Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. J Immunol 152:163–175

Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S 1999 SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50:213–219

Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V 1999 Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. J Med Chem 42:4650–4658

Schönbach C, Koh JL, Sheng X, Wong L, Brusic V 2000 FIMM, a database of functional molecular immunology. Nucleic Acids Res 28:222–224

Sette A, Buus S, Appella E et al 1989 Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. Proc Natl Acad Sci USA 86:3296–3300

Sette A, Vitiello A, Reherman B et al 1994 The relationship between class I binding affinity and immunogenicity of potential cytotoxic T-cell epitopes. J Immunol 153:5586–5592

Thornton JM, Edwards MS, Taylor WR, Barlow DJ 1986 Location of 'continuous' antigenic determinants in the protruding regions of proteins. EMBO J 5:409–413

DISCUSSION

Rammensee: I have a practical question. If you look at the affinity of the known T cell epitopes and line them up in order of highest to lowest IC_{50}s, do you see anything interesting? For instance, are the virus peptides on top and tumour peptides at the bottom?

Flower: We haven't done this. There is a lot of potential for data mining here, but you need both the resources and an appropriate motivation to undertake the task. I would expect that as our database grows we, or others, will do so.

Rammensee: The reason for this particular comparison is that it is thought that tumour epitopes in general have lower affinity for MHC than virus epitopes. It would be nice to see this result from your unbiased measurements.

Borras-Cuesta: Up to now we have been talking about predictions. We have not taken into account what characteristics peptides should have to induce certain responses. That is, it is known by experiments that a helper peptide that induces mainly interferons — that is Th1 cytokines — will drive towards cytotoxic

responses. However, if it is more Th0- or Th2-like, it will drive towards inducing antibodies. This is very important. When we predict peptides, we only predict that they bind. We don't predict that they will achieve their goal. As far as prediction is concerned, there are two things that need to be taken into account. One is binding to MHC. The other one — and amazingly we have not taken this seriously — is recognition by the TCR. What you mentioned about the epitopes in cancer is correct. Because of clonal deletion in cancer, the T lymphocytes that recognize the good epitopes are deleted. It is not just that they have to be bad binders; they can be good binders but poorly recognized, because the deletion is made at that end. All these people who are so good at predicting should take this seriously and predict binding to the TCR. It sounds very odd but it can be approximated. Certain amino acids do tend to have stronger interactions, such as the charged amino acids. If one considers the positions of proteins that point to the TCR, if they have these amino acid residues, they might be good candidates for T cell recognition. I suggest that this should be taken into account. I have written programs that consider this, but I don't have a good database to test this. The idea, however, is a logical one.

DeLisi: We have done that. I published a paper in 1996 looking at the whole ternary complex (Vasmatzis et al 1996), before the ternary complex was done experimentally. There are enough ternary complexes now so that by homologous extension we could probably get pretty good results.

Borras-Cuesta: It is amazing that we talk about prediction without considering this. If you use a helper peptide that produces interferon and induces a CTL response, this protects mice from a certain type of cancer. If you immunize with a peptide that induces response that is Th0 type, it protects partially. If you mix the helper peptide producing high interferon with the other one which is Th0, then you are back to square one and you get the same protection as Th0 alone.

Kellam: Is this saying that it is the strength of signalling through the TCR receptor?

Borras-Cuesta: It is the total effect. It is not just recognition by the TCR. If there is a lot of peptide presented by MHC, plus a good recognition, that is the best you can have. It is a chemical problem.

Kellam: And the local cytokine and signalling environment are important.

Borras-Cuesta: This will probably drive strong interactions. Many people think that strong interactions will drive towards production of interferon and a Th1-like response.

Rammensee: This will also depend on background genes. This is the case in the mouse: in BALB/c and B6 there is a strain-related difference in the Th1/Th2 balance. This needs to be taken into account. Darren Flower, with respect to the TCR interaction, will it be possible to predict the structure of all 10^{16} TCRs? For

each MHC peptide combination, can you work out which of the TCRs would bind? Is it just a computational problem?

Flower: If you could accurately predict the structures of a peptide–MHC–TCR complex, you had a big enough computer, and you could run your simulations for long enough, then I'm sure this would be achievable. However, the most important thing is to verify constantly what you do by undertaking directed experiments. You can't disentangle experiment from prediction. Experimental scientists will not believe you if you say your predictions work 100% of the time. You have to do at least some experiments in order to prove your predictions are accurate.

Gulukota: The lack of computational resources is a bottleneck, but an even bigger one is our understanding of the physical chemistry. Even given the most powerful computer we can imagine, we don't know enough about the interactions between the atoms.

DeLisi: When proteins fold or complexes form, polar and apolar groups often move between regions having very different dielectric properties, e.g. from being fully solvated in a denatured protein, to the interior of the native form. Such changes have substantial free energy components. What kind of salvation free energy functions did you do? Did you use explicit water?

Flower: Yes.

DeLisi: Is this why it has taken so long? Have you tried looking at some kind of semi-empirical calculation? We have developed several which are very effective and fast.

Flower: There are obviously many things that we have yet to try. We are still very much at the beginning of things.

DeLisi: You are in a position to be able to approach these problems, but you need to be able to speed up that part of the calculation. If you look into some effective free energy functions for solvation you might be able to do that. Now there are three ternary complexes available and you are in a position to start considering these things.

Flower: We are running a series of benchmarks at the moment so that we can believe that the results that we obtain from our simulations are essentially the same as the AMBER force field implemented in its native code. Once we are happy that the simulation itself is working correctly we will consider ways of taking this further and using some sort of solvation model.

Lybrand: A bigger issue here is one that you raised in your paper, and this relates to the quality of the affinity data you are comparing with. These are not easy experiments to do. Depending on the method used to measure affinities, or which lab does them, you get fairly different answers. When you are in the process of calibrating the various methods, which you are doing now, this is a source of great frustration.

Flower: This is why something like isothermal titration calorimetry (ITC) would be a much better way of obtaining data, but it takes a whole day to do a peptide and is very expensive. Speaking practically, it is not something one could use easily to measure hundreds of peptides. If you could take that as a gold standard, and compare the rest of your data with results from ITC, demonstrating good correlations, this would be a significant advance. I feel that the quality and reliability of the data is a limiting factor in studies such as these.

Lybrand: You will need these data, especially if you want to use a technique like Charles DeLisi implies, involving an empirical or semi-empirical model for representing the solvent in these systems. It is now well documented that if you put in all the explicit detail and do exhaustive simulations as you suggest — which by the way scale a lot better in AMBER than your chart showed — you can get quantitatively reproducible free energy predictions for large biomolecular complexes. One of the difficulties here is that in a system like a ternary complex, with the MHC peptide and TCR, the experimental numbers are often all over the place. Consequently, you don't know whether you are doing a good job or not. The technology is here to do these kinds of calculations. These system sizes are no longer particularly daunting. These are much smaller complexes than many people are now simulating. One of the difficulties that I see is that some of the experimental numbers we have here are still a bit messy, so we don't know for sure whether we are properly addressing issues in these systems or not.

Brusic: If the starting points and assumptions are not solid we have a problem. For example, it is often assumed that T cell epitopes are the only peptides that bind strongly to MHC molecules. Is this really the case?

Borras-Cuesta: Even if they don't bind all that well, they can still be immunogenic because the other side compensates. It is not only binding, but also recognition by the TCR that matters.

Brusic: We were discussing the QSAR analysis and other data-driven models. Neural networks can used for QSARs, so can other methods which may overlap. I consider that binding motifs, quantitative matrices, and artificial neural networks are essentially modelling the very same properties. They are just models that encode different levels of complexity. We can take a neural network that models peptide–MHC interactions and start reducing it. By removing hidden layers and use a linear activation function, our neural network would become a quantitative matrix. We can reduce this neural network further by removing connections from the architecture and get a neural network that is a binding motif. The key issue for modelling is matching the complexity of biological interaction with the complexity of the model that simulates that interaction. Currently we have a variety of computational models, and most of the models that we use are the simplification of the real process or system. Another important issue arising from Darren Flower's paper is on combining multiple methods. For example, we can do

quick-and-dirty methods for large-scale screening to identify promising targets. We can then do detailed analysis on the targets of interest using more complicated modelling methods.

DeLisi: With the TCR the mix may even be more complex, because the accessory molecules provide a lot of the stability for that interaction. In some cases they have to diffuse over. Basically we can view the TCR as just holding the complex together long enough for some accessory molecule to diffuse over. The dissociation time has to be just a little bit longer than the time constant for diffusion in the membrane in order for a stable complex to form. It is the ratio of those two rate constants — one diffusion in the membrane and the other dissociation — that may ultimately determine whether you get a stable complex. It is a biologically subtle situation. I don't think it is difficult to put those factors in, but it is a little more complicated than a simple biomolecular reaction.

Flower: If you are looking at whole cells interacting, an important phenomenon is formation of the immunological synapse, where all sorts of other accessory molecules are involved in the signal transduction. It is very complicated process and we are just modelling a very small part of it rather than the whole thing. There may be many other emergent properties of the system that we are completely ignoring, and these could be driving what is going on. If we could model the whole process this would make the prediction much more effective.

DeLisi: There is a subtlety. The reverse rate constant of the T cell receptor could be critically important. It sets the time scale for whether all those subsequent processes happen or not. You want to characterize that. It is not just a matter of the stability of the complex itself. If the rest of those steps take longer than the dissociation process you wouldn't get a high affinity. It is an interesting problem.

Lybrand: This is a nice example of where you can merge detailed molecular modelling with some of the higher-level mathematical models that we have talked about. Of course, we don't know all of the rate constants. I am not sure we either know all of the players involved in what is an incredibly complicated multimolecular complex — the functional T cell signalling apparatus. My experimental colleagues who are doing experiments to compare with our simulations now tell me that depending on differential rates you get dramatically different immune responses or read-outs. It gets even more complicated than just putting in simple binary or ternary rate constants. Again, this is an ideal situation where you can go from the detailed modelling to a more formal mathematical model for the entire assembly. This is the kind of thing we are trying to do in a number of different systems. You have to deal with what you have got in hand and can address in a more tangible way, because there are too many other unknown variables to address all these issues right now.

Flower: It would be nice to simulate two whole cells interacting, but we are some time away from this.

Lybrand: It may happen sooner than you think.

DeLisi: It depends what level you attempt this. If you get phenomenological at a different level it might be easier.

Perelson: I have a comment about the level we are at right now. Many of you are aware of the phenomenon of altered peptide ligands. You can have two peptides binding the same MHC, and the T cell will interact with both peptide–MHC complexes with the same free energy change, and yet one will stimulate and one won't. This effect depends on the reverse rate constants or equivalently the lifetime of the peptide–MHC–T cell receptor complex, as Charles DeLisi has pointed out. We need to know not only the energy of the reaction, but in terms of the functional response of the T cell the lifetimes of complexes are also very important. Many studies are involved with trying to optimise the affinity of interactions. When we start looking at functional responses at the level of the T cell it is not at all clear when one wants to have the highest possible affinity interaction. Typically if one is doing this one is restricting the number of T cells that will enter the response. There tend to be many fewer T cells that will respond at high affinity. We may want to look at breadth of response. One might also think about the same issue at the level of the MHC: to what extent do we want cross-reactivity of our peptides with a number of MHCs, rather than just designing very specifically for one MHC? For the functional response we have to worry about both the breadth and the strength of the response. When we start thinking about system-wide properties, we need to think of the whole T cell repertoire and how it responds.

Flower: The methods that others, such as Didier Rognan, and ourselves have been trying to develop will allow you to predict higher binders and lower binders, and explore the affinity range. We are now starting to gather together kinetic data so we can begin to model on rates and off rates in a similar way to the modelling of affinities.

Reference

Vasmatzis G, Cornette J, Sezerman U, DeLisi C 1996 TcR recognition of the MHC–peptide dimer: structural properties of a ternary complex. J Mol Biol 261:72–89

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

IMGT, the international ImMunoGeneTics information system®, http://imgt.cines.fr

Marie-Paule Lefranc

Université Montpellier II, Laboratoire d'ImmunoGénétique Moléculaire, LIGM, UPR CNRS 1142, Institut de Génétique Humaine, Montpellier, France

Abstract. IMGT, the international ImMunoGeneTics information system® (*http://imgt.cines.fr*), is a high quality integrated knowledge resource specializing in immunoglobulins (IG), T cell receptors (TR) and major histocompatibility complexes (MHC) and related proteins of the immune system (RPI) of human and other vertebrates, created in 1989 by LIGM at the Université Montpellier II, CNRS, Montpellier, France. IMGT provides a common access to standardized data which include nucleotide and protein sequences, oligonucleotide primers, gene maps, genetic polymorphisms, specificities, and 2D and 3D structures. IMGT includes five databases (IMGT/LIGM-DB, IMGT/3Dstructure-DB, IMGT/MHC-DB, IMGT/PRIMER-DB, IMGT/GENE-DB) Web resources ('IMGT Marie-Paule page') and interactive tools (IMGT/V-QUEST, IMGT/JunctionAnalysis, IMGT/PhyloGene, IMGT/LocusView, IMGT/GeneView, IMGT/GeneSearch, IMGT/StructureQuery). IMGT data are expertly annotated according to the rules of the IMGT Scientific chart based on IMGT-ONTOLOGY. IMGT tools are particularly useful for the analysis of the IG and TR repertoires in physiological normal and pathological situations. IMGT has important applications in medical research (autoimmune diseases, AIDS, leukaemias, lymphomas, myelomas), biotechnology related to antibody engineering (phage displays, combinatorial libraries) and therapeutic approaches (graft, immunotherapy). IMGT is freely available at *http://imgt.cines.fr*.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 126–142

The molecular synthesis and genetics of the immunoglobulin (IG) and T cell receptor (TR) chains is particularly complex and unique as it includes biological mechanisms such as DNA molecular rearrangements in multiple loci (three for IG and four for TR in humans) located on different chromosomes (four in humans), nucleotide deletions and insertions at the rearrangement junctions (or N-diversity), and somatic hypermutations in the IG loci (for review Lefranc & Lefranc 2001a,b). The number of potential protein forms of IG and TR is almost unlimited. Owing to the complexity and high number of published sequences, data

control, classification and detailed annotations are difficult tasks for the generalist databanks such as EMBL, GenBank, and DDBJ. These observations were the starting point of IMGT, the international ImMunoGeneTics information system® (*http://imgt.cines.fr*) (Lefranc 2001a) created in 1989 by the Laboratoire d'ImmunoGénétique Moléculaire (LIGM) at the Université Montpellier II, CNRS (Montpellier, France).

IMGT is a high quality integrated information system specializing in IG, TR, MHC and RPI of human and other vertebrates which consists of five databases (IMGT/LIGM-DB, IMGT/3Dstructure-DB, IMGT/MHC-DB, IMGT/PRIMER-DB, IMGT/GENE-DB), Web resources ('IMGT Marie-Paule page') and interactive tools (IMGT/V-QUEST, IMGT/JunctionAnalysis, IMGT/PhyloGene). IMGT expertly annotated data and tools are particularly useful for the analysis of the IG and TR repertoires in physiological and pathological situations. By its easy data distribution, IMGT has important implications in medical research (autoimmune diseases, AIDS, leukaemias, lymphomas, myelomas), biotechnology related to antibody engineering, (phage displays, combinatorial libraries) and therapeutic approaches (grafts, immunotherapy). IMGT is freely available at *http://imgt.cines.fr*.

IMGT databases

IMGT/LIGM-DB is a comprehensive database of IG and TR nucleotide sequences from human and other vertebrate species, with translation for fully annotated sequences, created in 1989 by LIGM and on the Web since July 1995 (Lefranc 2001a). In July 2003, IMGT/LIGM-DB contained 74 387 nucleotide sequences of IG and TR from 105 species.

IMGT/LIGM-DB data are provided with a user-friendly interface (Giudicelli et al 1997). The Web interface allows searches according to immunogenetic specific criteria and is easy to use without any knowledge of a computing language (Fig. 1). A selection is displayed at the top of the resulting sequences pages, so the users can check their own queries (Lefranc et al 1999). Users are able to modify their request or consult the results with a choice of nine possibilities (Lefranc 2002). IMGT/LIGM-DB data are also distributed by anonymous FTP servers at CINES (*ftp://ftp.cines.fr/IMGT/*), EBI (*ftp://ftp.ebi.ac.uk/pub/databases/imgt/*) and from many SRS (Sequence Retrieval System) sites. IMGT/LIGM-DB can be searched by BLAST or FASTA on different servers (EBI, IGH, INFOBIOGEN, Institut Pasteur, etc.).

IMGT/3Dstructure-DB is a database which provides the IMGT gene and allele identification and 2D graphical representations or Colliers de Perles of IG, TR, MHC and RPI with known 3D structures, created by LIGM, and on the Web sinc November 2001 (Ruiz & Lefranc 2002) (Fig. 2). In July 2003, IMGT/3Dstructure-DB contained 648 entries.

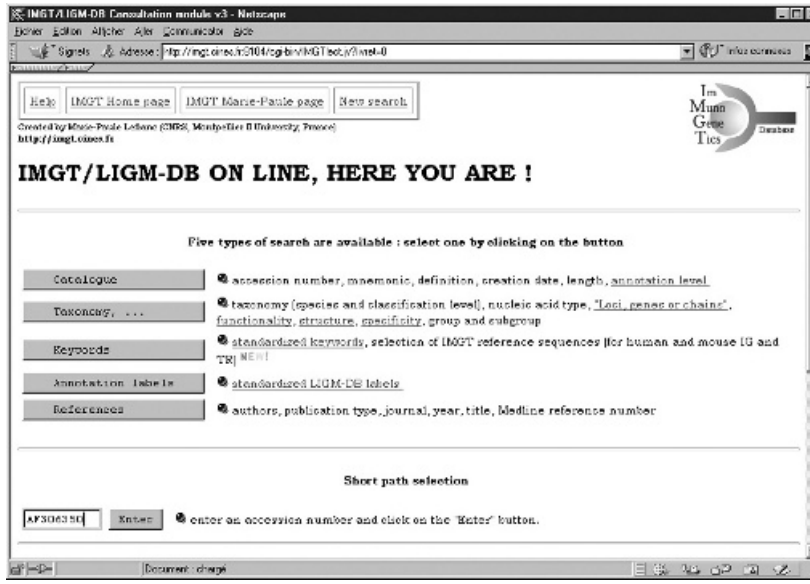


FIG. 1. IMGT/LIGM-DB (*http://imgt.cines.fr*) query interface.

IMGT/MHC-DB is hosted at EBI and comprises a database of the human MHC allele sequences, IMGT/MHC-HLA, developed by Cancer Research UK and Anthony Nolan Research Institute, London, UK, which has been on the Web since December 1998 (Robinson et al 2000) and a database of MHC class II sequences from non-human primates, IMGT/MHC-NHP, curated by BPRC, the Netherlands, on the Web since April 2002.

IMGT/PRIMER-DB is an oligonucleotide primer database for IG and TR-MHC, developed by LIGM and EUROGENTEC, Belgium.

IMGT/GENE-DB allows a search by gene name for IG and TR.

IMGT Web resources

IMGT Web resources ('IMGT Marie-Paule page') comprise 8000 HTML pages in the following sections: IMGT Scientific chart, IMGT Repertoire, IMGT Bloc-notes, IMGT Education, IMGT Aide-mémoire and IMGT Index.

IMGT scientific chart

The IMGT Scientific chart provides the controlled vocabulary and the annotation rules and concepts defined by IMGT for the identification, the description, the classification and the numerotation of the Ig and TCR data of human and other vertebrates (Giudicelli & Lefranc 1999).

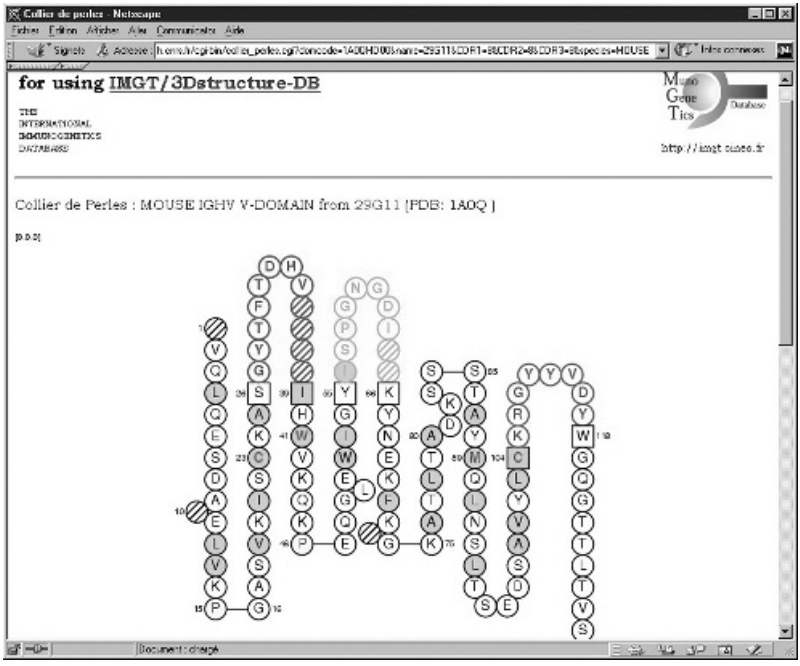


FIG. 2. IMGT/3Dstructure-DB (*http://imgt.cines.fr*). An example of Collier de Perles.

Concept of identification: standardized keywords. IMGT standardized keywords for IG and TR include general keywords, indispensable for the sequence assignments, and specific keywords, more specifically associated to particularities of the sequences or to diseases (Giudicelli et al 1997).

Concept of description: standardized sequence annotation. 387 feature labels are necessary to describe all structural and functional subregions that compose IG and TR sequences, whereas only seven of them are available in EMBL, GenBank or DDBJ. Annotation of sequences with these labels constitutes the main part of the expertise (Giudicelli et al 1997).

Concept of classification: standardized IG and TR gene nomenclature. The objective is to provide immunologists and geneticists with a standardized nomenclature per locus and per species which will allow extraction and comparison of data for the complex B and T cell antigen receptor molecules. The concepts of classification have been used to set up a unique nomenclature of human IG and TR genes, which was approved by the Human Genome Organization (HUGO) Nomenclature Committee (HGNC) in 1999 (Lefranc 2000a,b,c,d, Lefranc

2001b,c,d). The complete list of the human IG and TR gene names was entered by the IMGT Nomenclature Committee in GDB, Toronto, and LocusLink, NCBI, USA, and is available from the IMGT site (Lefranc & Lefranc 2001a,b). IMGT reference sequences have been defined for each allele of each gene based on one or, whenever possible, several of the following criteria: germline sequence, first sequence published, longest sequence and mapped sequence (Lefranc et al 1998). They are listed in the germline gene tables of the IMGT Repertoire (Pallarès et al 1998, 1999, Barbié & Lefranc 1998, Ruiz et al 1999, Folch & Lefranc 2000a,b, Scaviner & Lefranc 2000a,b). The protein displays show translated sequences of the alleles (*01) of the functional or ORF genes (Scaviner et al 1999, Folch et al 2000, Lefranc & Lefranc 2001a,b).

Concept of numerotation: the IMGT unique numbering. A uniform numbering system for IG and TR sequences of all species has been established to facilitate sequence comparison and cross-referencing between experiments from different laboratories whatever the antigen receptor (IG or TR), the chain type, or the species (Lefranc 1997, 1999). The IMGT unique numbering represents a big step forward in the analysis of the IG and TR sequences of all vertebrate species. It has allowed (i) a standardized description of the allele polymorphisms (Lefranc et al 1998, 1999) and of the IG somatic hypermutations, and (ii) the redefinition of the limits of the FR and CDR of the IG and TR variable domains (Lefranc et al 2003). The FR-IMGT and CDR-IMGT lengths become in themselves crucial information which characterize variable regions belonging to a group, a subgroup and/or a gene. Moreover, it gives insight into the structural configuration of the domains and opens interesting views on the evolution of these sequences, since this numbering has been applied with success to all the sequences belonging to the V-set and C-set of the immunoglobulin superfamily.

IMGT Repertoire

IMGT Repertoire is the global Web Resource in ImMunoGeneTics for the IG, TR, MHC and RPI of human and other vertebrates, based on the IMGT Scientific chart. IMGT Repertoire provides an easy-to-use interface to carefully and expertly annotated data on the genome, proteome, polymorphism and structural data of the IG, TR, MHC and RPI. Only titles of this large section are quoted here. **Genome** data include chromosomal localizations, locus representations, locus description, gene tables, lists of genes and links between IMGT, HUGO, GDB, LocusLink and OMIM, correspondence between nomenclatures. **Proteome** and **polymorphism** data are represented by protein displays, alignments of alleles, tables of alleles and allotypes. **Structural data** comprise Colliers de Perles, FR-IMGT and CDR-IMGT lengths, and 3D representations (Ruiz et al 2000, Ruiz & Lefranc 2002).

Other IMGT Web sections

The IMGT Bloc-notes provides numerous hyperlinks towards the Web servers specializing in immunology, genetics, molecular biology and bioinformatics (Lefranc 2000e). IMGT Education and IMGT Aide-mémoire provide useful information for students (figures, tutorials). IMGT Index is a fast way to access data when information has to be retrieved from different parts of the IMGT site.

IMGT interactive tools

IMGT/V-QUEST

IMGT/V-QUEST (V-QUEry and STandardization) is an integrated software tool for IG and TR. This easy-to-use tool analyses an input IG or TR germline or rearranged variable nucleotide sequences (Fig. 3). IMGT/V-QUEST results comprise the identification of the V, D and J genes and alleles and the nucleotide alignments, by comparison with sequences from the IMGT reference directory, the delimitations of the FR-IMGT and CDR-IMGT based on the IMGT unique numbering, the protein translation of the input sequence, the identification of the JUNCTION and the V-REGION Collier de Perles. The set of sequences from the

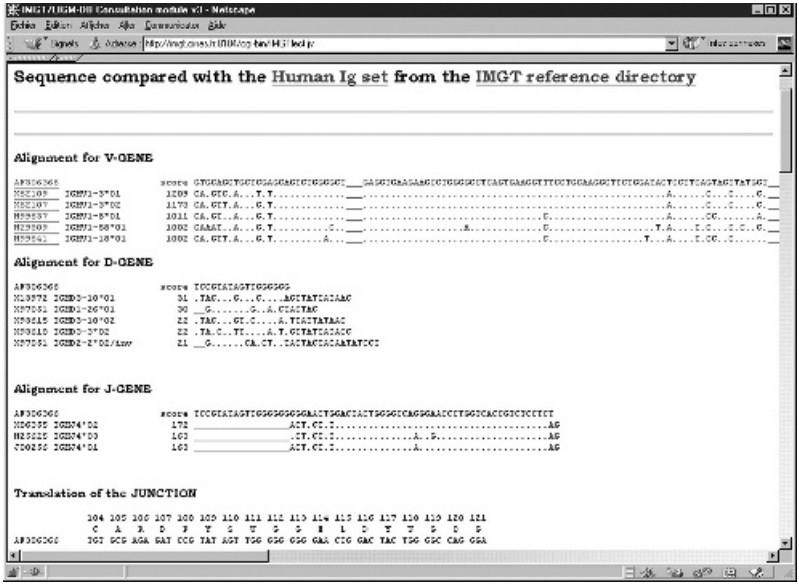


FIG. 3. IMGT/V-QUEST (*http://imgt.cines.fr*) results. IMGT/V-QUEST compares the input germline or rearranged IG or TR variable sequences with the IMGT/V-QUEST reference directory sets. The IMGT/V-QUEST results comprise the translation of the JUNCTION for rearranged sequences, and also, not shown in the figure, the delimitations of the FR-IMGT and CDR-IMGT, the protein translation and the V-REGION Collier de Perles.

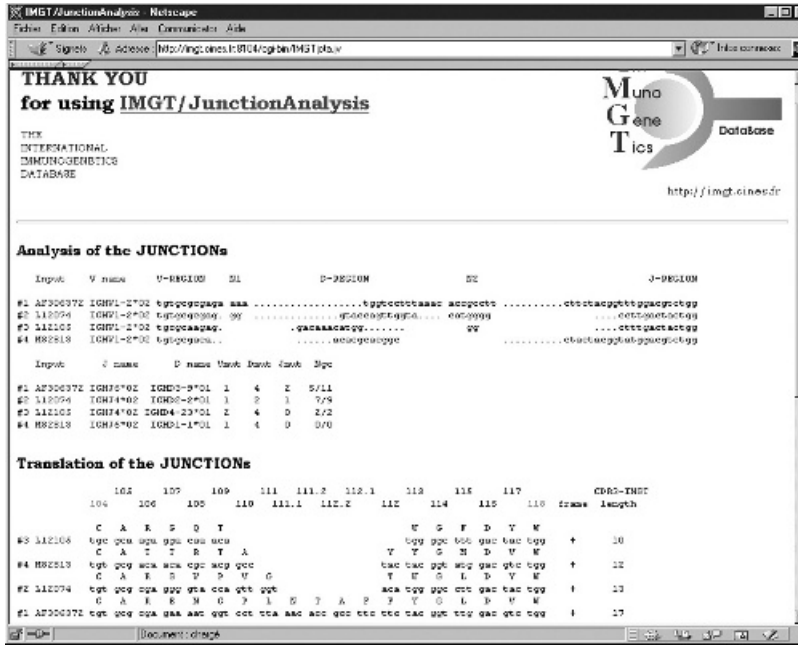


FIG. 4. IMGT/JunctionAnalysis (*http://imgt.cines.fr*) results. The IMGT/JunctionAnalysis results comprise, for each junction, the identification of the D-GENE and allele, the identification of the P and N regions (N1, N2, etc.) and their precise delimitations, and the junction translation. The CDR3-IMGT numbering is according to the IMGT unique numbering for V-DOMAIN. Vmut, Dmut and Jmut correspond to the number of mutations in the input junction sequence by comparison to the germline allele sequences. Ngc is the ratio of the number of g+c nucleotides to the total number of nucleotides in the N regions. IMGT/JunctionAnalysis analyses, in a single search, an unlimited number of junctions provided that the V-GENE and J-GENE allele IMGT names are identified.

IMGT reference directory, used for IMGT/V-QUEST, can be downloaded in FASTA format from the IMGT site.

IMGT/JunctionAnalysis

IMGT/JunctionAnalysis is a tool, complementary to IMGT/V-QUEST, which provides a thorough analysis of the V–J and V–D–J junctions of IG and TR rearranged genes (Fig. 4). IMGT/JunctionAnalysis identifies the D-GENE and allele involved in the IGH, TRB and TRD V–D–J rearrangements by comparison with the IMGT reference directory, and delimits precisely the P, N and D regions. Several hundred junction sequences can be analysed simultaneously.

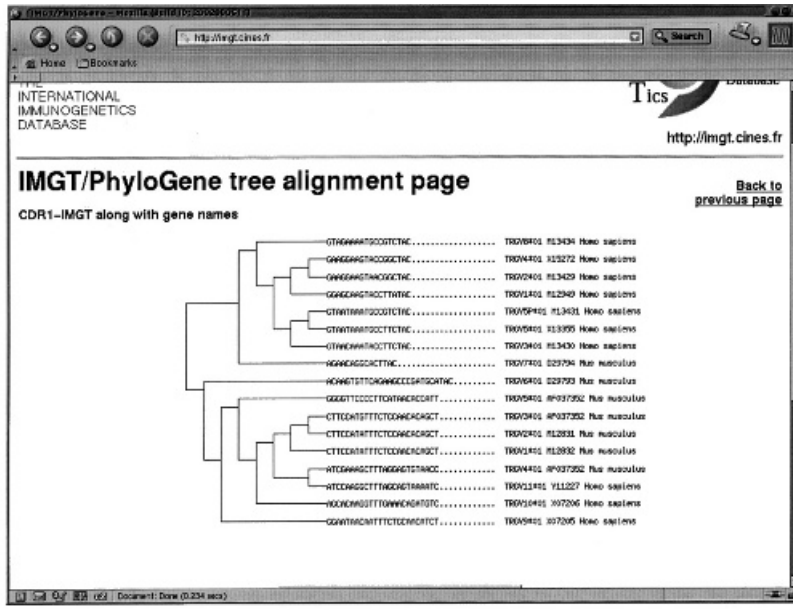


FIG. 5. IMGT/PhyloGene (*http://imgt.cines.fr*) resulting phylogenetic tree for the human and mouse TRGV genes.

Other IMGT tools

IMGT/PhyloGene is an online package for comparative analysis of IG and TR sequences (Fig. 5). IMGT/GeneSearch, IMGT/GeneView and IMGT/LocusView provide a display of physical maps. IMGT/StructureQuery is a tool for 3D structure analysis of the IG, TR, MHC and RPI.

IMGT-ONTOLOGY and IMGT interoperability

IMGT-ONTOLOGY

IMGT distributes high quality data with an important incremental value added by the IMGT expert annotations, according to the rules described in the IMGT Scientific chart. IMGT has developed a formal specification of the terms to be used in the domain of immunogenetics and bioinformatics to ensure accuracy, consistency and coherence in IMGT. This has been the basis of IMGT-ONTOLOGY (Giudicelli & Lefranc 1999), the first ontology in the domain, which allows the management of the immunogenetics knowledge for all vertebrate species. Control of coherence in IMGT combines data integrity control and biological data evaluation (Giudicelli et al 1998a,b).

IMGT interoperability

Since July 1995, IMGT has been available on the Web at *http://imgt.cines.fr*. IMGT provides biologists with an easy to use and friendly interface. Since January 2000, the IMGT WWW Server at Montpellier was accessed by more than 210 000 sites. IMGT has an exceptional response with more than 120 000 requests a month. Two-thirds of the visitors are equally distributed between the European Union and the USA.

Conclusion

The information provided by IMGT is of much value to clinicians and biological scientists in general (Lefranc 2002, 2003). IMGT is designed to allow a common access to all immunogenetics data, and a particular attention is given to the establishment of cross-referencing links to other databases pertinent to the users of IMGT.

Citing IMGT

Authors who make use of the information provided by IMGT should cite Lefranc (2001a) as a general reference for the access to and content of IMGT, and quote the IMGT home page URL, *http://imgt.cines.fr*.

Acknowledgements

IMGT is funded by the European Union's 5th PCRDT programme (QLG2-2000-01287), the Centre National de la Recherche Scientifique (CNRS) and the Ministère de l'Education Nationale et de la Recherche.

References

Barbié V, Lefranc M-P 1998 The human immunoglobulin kappa variable (IGKV) genes and joining (IGKJ) segments. Exp Clin Immunogenet 15:171–183

Folch G, Lefranc M-P 2000a The human T cell receptor beta variable (TRBV) genes. Exp Clin Immunogenet 17:42–54

Folch G, Lefranc M-P 2000b The human T cell receptor beta diversity (TRBD) and beta joining (TRBJ) genes. Exp Clin Immunogenet 17:107–114

Folch G, Scaviner D, Contet V, Lefranc M-P 2000 Protein displays of the human T cell receptor alpha, beta, gamma and delta variable and joining regions. Exp Clin Immunogenet 17: 205–215

Giudicelli V, Lefranc M-P 1999 Ontology for immunogenetics: IMGT-ONTOLOGY. Bioinformatics 12:1047–1054

Giudicelli V, Chaume D, Bodmer J et al 1997 IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 25:206–211

Giudicelli V, Chaume D, Lefranc M-P 1998a IMGT/LIGM-DB: a systematized approach for ImMunoGeneTics database coherence and data distribution improvement. In: Glasgow G, Littlejohn T, Major F, Lathrop R, Sankoff D, Sensen C (eds) Proceedings of the Sixth

International Conference on Intelligent Systems for Molecular Biology (ISBM-98), Montreal, June–July 1998, AAAI Press, p 59–68
Giudicelli V, Chaume D, Mennessier G et al 1998b IMGT, the international ImMunoGeneTics database: a new design for immunogenetics data access. In: Cesnik B, McCray AT, Scherrer JR (eds) Proceedings of the Ninth World Congress on Medical Informatics (MEDINFO' 98), Seoul 1998, IOS Press, Amsterdam, p 351–355
Lefranc M-P 1997 Unique database numbering system for immunogenetic analysis. Immunol Today 18:509
Lefranc M-P 1999 The IMGT unique numbering for immunoglobulins, T cell receptors and Ig-like domains. The Immunologist 7:132–136
Lefranc M-P 2000a Nomenclature of the human immunoglobulin genes. Current protocols in immunology. John Wiley & Sons, New York, USA (suppl 40) A.1P.1–A.1P.37
Lefranc M-P 2000b Nomenclature of the human T cell receptor genes. Current protocols in immunology. John Wiley & Sons, New York, USA (suppl 40) A.1O.1–A.1O.23
Lefranc M-P 2000c Locus maps and genomic repertoire of the human Ig genes. The Immunologist 8:80–87
Lefranc M-P 2000d Locus maps and genomic repertoire of the human T-cell receptor genes. The Immunologist 8:72–79
Lefranc M-P 2000e Web sites of interest to immunologists. Current protocols in immunology. John Wiley & Sons, New York, USA, A.1J.1–A.1J.33
Lefranc M-P 2001a IMGT, the international ImMunoGeneTics database. Nucl Acids Res 29:207–209
Lefranc M-P 2001b Nomenclature of the human immunoglobulin heavy (IGH) genes. Exp Clin Immunogenet 18:100–116
Lefranc M-P 2001c Nomenclature of the human immunoglobulin kappa (IGK) genes. Exp Clin Immunogenet 18:161–174
Lefranc M-P 2001d Nomenclature of the human immunoglobulin lambda (IGL) genes. Exp Clin Immunogenet 18:242–254
Lefranc M-P 2002 IMGT, the international ImMunoGeneTics database: a high-quality information system for comparative immunogenetics and immunology. Dev Comp Immunol 26:697–705
Lefranc, M.-P 2003 IMGT databases, web resources and tools for immunoglobulin and T cell receptor sequence analysis, *http://imgt.cines.fr*. Leukemia 17:260–266
Lefranc M-P, Lefranc G, 2001a The Immunoglobulin FactsBook. Academic Press, London, UK
Lefranc M-P, Lefranc G 2001b The T cell receptor FactsBook. Academic Press, London, UK
Lefranc M-P, Giudicelli V, Busin C et al 1998 IMGT, the international ImMunoGeneTics database. Nucl Acids Res 26:297–303
Lefranc M-P, Giudicelli V, Ginestoux C et al 1999 IMGT, the international ImMunoGeneTics database. Nucl Acids Res 27:209–212
Lefranc M-P, Pommié C, Ruiz M et al 2003 IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. Dev Comp Immunol 27:55–77
Pallarès N, Frippiat JP, Giudicelli V, Lefranc M-P 1998 The human immunoglobulin lambda variable (IGLV) genes and joining (IGLJ) segments. Exp Clin Immunogenet 15:8–18
Pallarès N, Lefebvre S, Contet V, Matsuda F, Lefranc M-P 1999 The human immunoglobulin heavy variable (IGHV) genes. Exp Clin Immunogenet 16:36–60
Robinson J, Malik A, Parham P, Bodmer JG, Marsh SGE 2000 IMGT/HLA database—a sequence database for the human major histocompatibility complex. Tissue Antigens 55:280–287
Ruiz M, Lefranc M-P 2002 IMGT gene identification and Colliers de Perles of human immunoglobulin with known 3D structures. Immunogenetics 53:857–883

Ruiz M, Pallarès N, Contet V, Barbié V, Lefranc M-P 1999 The human immunoglobulin heavy diversity (IGHD) and joining (IGHJ) segments. Exp Clin Immunogenet 16:173–184

Ruiz M, Giudicelli V, Ginestoux C et al 2000 IMGT, the international ImMunoGeneTics database. Nucl Acids Res 28:219–221

Scaviner D, Lefranc M-P 2000a The human T cell receptor alpha variable (TRAV) genes. Exp Clin Immunogenet 17:83–96

Scaviner D, Lefranc MP 2000b The human T cell receptor alpha joining (TRAJ) genes. Exp Clin Immunogenet 17:97–106

Scaviner D, Barbié V, Ruiz M, Lefranc M-P 1999 Protein displays of the human immunoglobulin heavy, kappa and lambda variable and joining regions. Exp Clin Immunogenet 16:234–240

DISCUSSION

Flower: I have not used the IMGT database much myself, but colleagues who are computationally oriented and also various lab immunologists have said to me that while this database seems to have a fantastic amount of information, which is potentially very useful, what they find difficult is actually searching it. It appears to be quite difficult to extract specific information in some cases. Are you planning to improve the querying functions?

Lefranc: IMGT (*http://imgt.cines.fr*) is an integrated information system specializing in IG, TR, MHC and RPI, and as such, includes several databases (IMGT/LIGM-DB, IMGT/PRIMER-DB, IMGT/GENE-DB, IMGT/3Dstructure-DB, etc.), Web resources ('IMGT Marie-Paule page') comprising 8000 HTML pages, and several interactive tools. Queries by users can therefore be very diverse. Navigating through the Web resources and using the IMGT tools is quite easy. Searching the databases is a little more sophisticated because it requires that the users are aware of the level of knowledge available in the literature in immunogenetics to make the right query. IMGT/LIGM-DB, the first and largest IMGT database, comprises 75 000 nucleotide sequences of IG and TR from human and 104 other vertebrate species. IMGT/LIGM-DB data are provided with a user-friendly interface (*http://imgt.cines.fr*). The Web interface allows searches according to immunogenetic specific criteria and is easy to use without any knowledge of a computing language. Selection is displayed at the top of the resulting pages, so the users can check their own queries. Users have the possibility to modify their request or consult the results. They can (1) add new conditions to increase or decrease the number of resulting sequences, (2) view details concerning the selected sequences and choose among nine possibilities: annotations, IMGT flat file, coding regions with protein translation, catalogue and external references, sequence in dump format, sequence in FASTA format, sequence with three reading frames, EMBL flat file, IMGT/V-QUEST (with automatically generated Collier de Perles), or (3) search for sequence fragments corresponding to a

particular label. IMGT/LIGM-DB data are also distributed by anonymous FTP servers at CINES (*ftp://ftp.cines.fr/IMGT/*) and EBI (*ftp://ftp.ebi.ac.uk/pub/databases/imgt/*) and from many SRS (sequence retrieval system) sites. IMGT/LIGM-DB can be searched by BLAST or FASTA on different servers (EBI, IGH, INFOBIOGEN, Institut Pasteur, etc.).

Margalit: Do you curate the data?

Lefranc: I have six people working full time on the IMGT/LIGM-DB annotations. We daily receive data from EMBL (and, via EMBL, from GenBank and DDBJ). These data (about 500 sequences a week) are checked, standardized IMGT keywords are added and data are entered in the IMGT/LIGM-DB database (at the 'keyword annotation' level). Sequences are then annotated by batch. All the corresponding tables of IMGT Repertoire and the IMGT tools are updated and, if necessary, new tables and HTML pages are created on specific subjects. This work is carried out by eight people. When you query the IMGT/LIGM-DB database you can tell what level of annotation has been provided: 'keyword annotation level' or 'fully annotated' and, if so, whether it is 'automatically annotated' or whether it is 'annotated by annotators'. 'Automatically annotated' in IMGT/LIGM-DB means a level of sequence annotation as high as the one done by the annotators, since the tools used for the automatic annotations are developed based on the IMGT Scientific chart rules. However, this only applies to cDNAs from species (human and mouse) for which all IG and TR genes are known. Genomic sequences from gene clusters and sequences from other species can only be annotated by the annotators. Comments based on the literature search are always added manually.

Brusic: Your database is one of the highest quality biological databases. It is based on serious classification, ontology and modelling work. Can you tell us about your struggle for database quality and acceptance by researchers?

Lefranc: It took 10 years to have the immunoglobulin and T cell receptor genes widely accepted in the genome databases. Indeed, there was a general fear, at the end of the 1980s, that the IG and TR genes would introduce a bias in the genome databases by their number. Moreover the generalist databases did not know how to deal with the molecular synthesis and genetics of the IG and TR chains which is particularly complex and unique since it includes DNA molecular rearrangements, nucleotide deletions and insertions at the rearrangement junctions, and hypermutations in the Ig loci. It was, in 1989, at the 10th International Human Gene Mapping Workshop (HGM 10) in New Haven, that for the first time the human TRG genes on chromosome 7 that we sequenced in 1984–1985 were entered in the Genome Database (GDB). During that meeting, we demonstrated that the standards set up, in our lab, for the identification, the classification and the description of the human TRG genes and alleles could be applied to all genes and alleles of the other IG and TR. This led to the creation in June 1989, at HGM 10, of IMGT. IMGT is now the international reference in immunogenetics, and had, for

its 10th anniversary, the honour of the Nucleic Acids Research (NAR) cover of the January database issue. IMGT data annotations for IG and TR of all vertebrate species are based on the 'IDENTIFICATION', 'CLASSIFICATION', 'DESCRIPTION', 'NUMEROTATION' and 'OBTENTION' concepts of IMGT-ONTOLOGY (Giudicelli & Lefranc 1999), the first ontology in immunogenetics and in immunoinformatics. The rules based on the IMGT-ONTOLOGY concepts are described in the IMGT Scientific chart (*http://imgt.cines.fr*). In 1999, IMGT gene names and definitions for all the human Ig and TCR genes, based on the IMGT-ONTOLOGY 'CLASSIFICATION' concept, were approved by the Human Genome Organization Gene Nomenclature Committee (HGNC) (Lefranc & Lefranc 2001a,b). The IMGT Nomenclature Committee is delegated by HGNC to assign new IG and TR gene symbols and alleles, via the IMGT/LIGM-DB database (Wain et al 2002).

IMGT combines data integrity control, biological evaluation and interoperability with other databases. IMGT/LIGM-DB (75 000 entries of IG and TR from vertebrates) is complementary to the generalist database SWISS-PROT (120 000 entries from bacteria, viruses, plants, invertebrates and vertebrates). Only a few representatives of IG and TR are present in SWISS-PROT in order to avoid the bias of introducing a large amount of specialized data in a generalist database. The reciprocity with the genome databases works quite nicely. However we have to be aware that data are harvested automatically in the generalist genome databases. At the beginning the data in LocusLink were very clean with just one IMGT reference sequence entered manually for each IG and TR gene. Now it isn't possible to recognize which one is the IMGT reference sequence. GDB has been much more cautious and had identified the IMGT reference sequence as Seq.@IMGT: For us the concept of reference sequence is important. The IMGT reference sequences for IG and TR have been defined based on one or, whenever possible, several of the following criteria: germline sequence, first sequence published, longest sequence, mapped sequence. Interestingly, most of the human IG and TR genes had been sequenced before the human genome project. When the complete genome was published we already had identified most of the reference sequences. We are confident that the oldest sequences are usually the best ones, because people were sequencing manually, on both strands, and took a year or so for each gene. We also check carefully that the reference sequences have been mapped in a physical way (on a phage, cosmid, YAC, etc.) and not only obtained by PCR from genomic DNA.

Littlejohn: Turning to the technological issues, one of the things that impresses me about the ENSEMBL database is that it allows multiple options for accessing the data: I can download the data, I can use their WWW interface or I can use a direct MySQL interface to the data. If I understand your technology, you are not using relational databases at the back end — is this correct?

Lefranc: IMGT is an information system which includes several relational databases, 8000 HTML pages and several tools. The three largest databases (IMGT/LIGM-DB, IMGT/GENE-DB and IMGT/PROTEIN-DB) use Sybase. Two databases (IMGT/3Dstructure-DB and IMGT/PRIMER-DB) use MySQL.

Littlejohn: Do you allow connections to the Sybase database directly? This is the first technological issue to get around the problems that people have seen with the flexibility of querying. If you have a direct access to the database using ODBC to the database, then this allows tremendous flexibility in querying.

Lefranc: We have built an Application Programming Interface (API) to access the database and its software tools and to facilitate the integration of IMGT data into applications developed by other laboratories. The information for API direct links to IMGT knowledge data, sequence data and human gene data is provided in the IMGT Informatics page (*http://imgt.cines.fr/informatics/*).

Littlejohn: In this case there are no theoretical problems with the complexity of queries that you allow. If I could make a secondary comment, what I don't like about ENSEMBL is the poor description of the metadata in the databank. The only way to do this is to deduce it from the table structure and field names. My feeling from IMGT is that you have better descriptions of metadata, which is ideal.

Brusic: The most appropriate data model will depend on the purpose of the database. If we want to use a database for data extraction and a quick search the relational model is very useful. However, for a higher-level analysis relational model it is not adequate. The issue here is how to provide a solution for problems based on different and often contradictory requirements. We need to discuss the database issues in more detail.

Littlejohn: I am not a huge advocate of relational databases, as biological data are object-oriented not tabular, but at the end of the day you have got to go with technologies that have been adopted broadly by the community. In some ways it is not a bad *de facto* standard, at least as a raw storage mechanism.

Brusic: We should not judge a particular database model as bad or good but determine whether its strengths correlate with our requirements.

Littlejohn: Most of the immunoinformatics we are discussing will draw on databases of one form or another. One of the great problems with databases is their upkeep and curation. We have a good example here of high quality curation, which requires six full time curators. SWISS-PROT has some 20 staff full time managing their databank.

Lefranc: Six people are indeed annotating the IMGT/LIGM-DB data, but our total lab consists of some 20 people. Eight people work on the IMGT Repertoire, tools updates and the other IMGT databases (IMGT/GENE-DB, IMGT/PROTEIN-DB, IMGT/PRIMER-DB, IMGT/3Dstructure-DB). Two people work on the interface and computing development. I am very keen that there is

active immunoinformatics research behind IMGT, so we have three or four PhD students working alongside the databases.

Littlejohn: It seems that most funding agencies are not particularly interested in funding the upkeep of what are essential data.

De Groot: NIH is very interested: they have developed an epitope mapping Request for Applications (RFP).

Flower: But they want to create a group of 20 people to run a database that will presumably subsume everyone else's. It is not necessarily a collaborative exercise.

De Groot: It could be a collaborative exercise.

Brusic: Biological databases don't usually make money. Rather, they represent a basic resource. However, when developing a database I always keep on mind what are the important applications of such a data set. A good application justifies the existence of the database and the effort for developing it.

Littlejohn: The problem is not making money, or even usability or utility. It is the currency of science — publication in high-quality peer-reviewed journals. It has typically been difficult to get a high-quality publications based on this kind of work. Nevertheless, it is a critical resource.

Brusic: A publication of a database is a matter of making it good, interesting, or new. People who try to reproduce what is already around will have a hard time trying to publish a database paper.

Flower: It is really the scale that you look at. Within immunoinformatics you might say that our database is different to that of other people because it has different kinds of data within it. But someone from outside the field will say it is exactly the same as other immunoinformatic databases.

Marsh: To answer this question, the first funding for the IMGT/HLA Sequence Database came through the grants we put together with Marie-Paule Lefranc that were EU funded. In 2000 we stopped receiving any EU money for that project. It had always been my aim to get sponsorship from commercial users. There are many people out there creating HLA typing reagents. These companies wanted a good data set to show the regulatory bodies. They know they couldn't do it themselves. We now have a situation where these companies and a number of other organization put money into a central pot that funds the work we do. It is tenuous in that we only ever know that we are going to get 12 months funding in advance, but it does seem to work.

Littlejohn: The difficulty in this case would be if your database were seen to have no direct relevance to a commercial partner. I wonder how well SWISS-PROT is doing with a similar model. The community is even seeing a movement away from publishing in journals because many scientists believe that the knowledge they create in the lab is for the public good should be accessible to all people at zero cost. But I'm wondering whether a journal-style model might be a better way for databases.

Flower: You cannot rely on commercial funding for databases, because the commercial realities of individual companies might change very rapidly. We want — we need — long-term stability.

Littlejohn: I think it has to be seen as infrastructure and paid for as such.

Flower: We need flesh and blood people to look at the information that is being mined. This makes databases expensive. You cannot rely solely on electronic data capture or text mining.

Marsh: It is relatively easy to source money to create a database, but it is more difficult to find support to maintain the database once it is up and running.

Bernaschi: I find it strange that there is so much discussion of the importance of databases and their quality, yet there is no agreement on database development and maintenance. I see two possible choices. One, there is a company that makes a business out of databases, selling access to reliable reference databases. This is what has happened in the financial markets. There is lots of information, much of which is not reliable. Then there are two or three companies who make a lot of money building financial market databases that people really use. There are other databases that are available for free, but people working seriously in financial markets always use Bloomberg or Datastream. In biology it would not be easy to find companies willing to do this. So we are left with the second option. If we really believe that databases are important a major effort should be made to find agreement among database developers. It is a waste of time looking through multiple databases trying to sort out which information is reliable. And with regard to clever access to the data, many people work in the field of data mining, so why should we re-invent the wheel? We should talk with other people to implement strategies for data mining in this field.

Littlejohn: Developing viable business models for companies curating and selling data in life sciences is very hard, because biological research is not like finance. There are only around 20 large pharmaceutical companies globally that will pay lots of money for data, and thousands of smaller biotech companies that may pay something for data, but these numbers are tiny in comparison with, for example, the number of banks and financial analysts. I have seen a number of bioinformatics companies fail in this area, or be forced to change direction. As an example, Celera is a company that has changed its business model from an information company to a pharmaceutical company. Information is a tough business in life sciences. For data mining, there is a lot of transplantation of technology and expertise from other industries.

Brusic: Life sciences are fragmented. To obtain the expertise for a particular subfield, first we need to employ an expert. Second we need to train other members of the team so that they can talk to the expert. A huge diversity within life sciences prevents us from finding the 'big answers'.

Littlejohn: I have been attending bioinformatics meetings for 10 years, but this is the first that has been specific, focusing on just immunoinformatics. This is fantastic, but it reflects the specificity of the domain knowledge required to come to terms with a particular informatics problem.

References

Giudicelli V, Lefranc MP 1999 Ontology for immunogenetics: the IMGT-ONTOLOGY. Bioinformatics 15:1047–1054

Lefranc MP, Lefranc G 2001a The immunoglobulin FactsBook. Academic Press, London

Lefranc MP, Lefranc G 2001b The T cell receptor FactsBook. Academic Press, London

Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S 2002 Guidelines for human gene nomenclature. Genomics 79:464–470

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

Generating data for databases — the peptide repertoire of HLA molecules

Stefan Stevanović, Claudia Lemmel, Maik Häntschel and Ute Eberle

Eberhard-Karls-Universität Tübingen, Institut für Zellbiologie, Abteilung Immunologie, Auf der Morgenstelle 15, D-72076 Tübingen, Germany

Abstract. During the past few years, a huge amount of information about HLA-presented peptides has been compiled: several thousand naturally processed ligands of such cell surface receptors are already known. Nevertheless, our knowledge covers only a minute proportion of the total peptide repertoire. The overall amount of different peptides presented by one given HLA class I molecule lies between 1000 and 10 000 individual sequences per cell. There is, however, no HLA molecule of which more than 100 ligands have been published so far. The situation is further complicated by the fact that different cells present different sets of peptides by the same HLA molecules, a feature that provides great hope for immunotherapy. We have been analysing HLA-presented peptides for many years for three reasons. First, the basic rules of peptide presentation (the 'peptide motifs') had to be established. Second, the listing of individual peptides presented by HLA molecules is steadily continuing, although a comprehensive catalogue of all possible HLA-presented peptides is utopical in our days. Third, quantitative differences in the presentation of individual HLA ligands provide information about the dynamic state of the host cells. Comprehensive information about HLA-presented peptides enables accurate epitope prediction and provides a basis for diagnostic assessment and therapeutic intervention.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 143–164

After several years of HLA ligand analysis, we introduced a listing of MHC ligands and peptide motifs (Rammensee et al 1995) as a comprehensive source of information. Very soon after, the number of identified sequences became too large for a printed listing and grew too fast, so we established the internet database SYFPEITHI (*www.syfpeithi.de*, Rammensee et al 1999). This bioinformatic tool provides information about thousands of MHC-presented peptides — immunogenic or not — and offers T-cell epitope prediction for a number of MHC molecules as a service.

The immunologist's interest in HLA-presented peptides extends from basic research to clinical application: The interaction between MHC specificity pockets (Garrett et al 1989) and bound ligands is as interesting a feature as the eradication of tumours by T cell recognition of MHC-presented peptides (Zwaveling et al 2002).

Not only immunogenic HLA-peptide complexes may turn out as important for diagnostic and therapeutic purposes: ligands derived from housekeeping proteins may indicate the normal state of a cell (at least with respect to antigen presentation). Viral or bacterial infections or even malignant transformations may be recognized during a very early stage by ligands from disease-associated proteins even if no immune reaction can be monitored. For a detailed immunoanalysis, the density of certain HLA ligands can be investigated either in absolute numbers or as a ratio between different cells.

Usually, HLA ligands are isolated by immunoprecipitation of MHC molecules, followed by acid-mediated peptide release (Rötzschke et al 1990). This strategy has been performed in many variations, and usually ends up with a low molecular weight fraction of peptides which are then separated by high performance liquid chromatography (HPLC). One problem in HLA ligand analysis is the small overall amount of peptides: from 10 billion HLA-expressing cells, less than one microgram of HLA ligand can be retrieved. The different abundance of individual peptides and the high complexity of the peptide pool represent major obstacles in obtaining fractions that contain pure individual sequences, which are present only in low nanogram quantities.

The classical method: analysis of peptide pools

In the early 1990s, it was the very low sensitivity of peptide analysis as well as poor recoveries of MHC-bound peptides that forced us to sequence complete peptide pools eluted from MHC molecules by Edman degradation (Falk et al 1991). Although in the meantime methods have been optimized and are now more sensitive by a factor of >1000, pool sequencing still represents a very quick and reliable method to gain comprehensive information about allele-specific peptide motifs. Pool data reveal positions and importance of anchor amino acids, show auxiliary anchors and preferred residues in every sequence position of MHC ligands, and may even give an impression of under-represented amino acids in certain positions. Thus, a basic peptide motif can be established by one analytical process that takes half a day — much less time than for analysis and alignment of more than 20 individual ligands. There are, however, some limitations intrinsic to the method: the N-terminal position is often hard to interpret because of high background noise, the C-terminal position because of fading signals due to sample loss. In addition, some amino acids escape the method (cysteine, in part also tryptophan) because of chemical instability, and finally, the absence of distinct amino acids in a given sequence position is difficult to recognize due to background effects.

Analysis of individual ligands by mass spectrometry

Although the first individual sequences of MHC ligands were defined by Edman degradation (Van Bleek & Nathenson 1990, Falk et al 1991) and many groups followed this strategy for several years (DiBrino et al 1993, Barber et al 1997), the limitations of this approach are obvious: only peptides that are well-separated from other sequences and highly dominant in their quantity can be analysed. Analysis of sequences which are less abundant and contaminated by other peptides leads to misinterpretations resulting in artificial sequences. In contrast, mass spectrometry, especially techniques using quadrupol instruments, can select single ion species from dozens of other peptides present in the same fraction. Therefore, mass spectrometry seems nowadays to represent the only technology that is able to analyse hundreds or even thousands of MHC-presented peptides from one biological source. From the pioneering work of Hunt and colleagues, we received a first estimation of the complexity of HLA ligands (Hunt et al 1992) and learned how to sequence peptides by tandem electrospray mass spectrometry. Later, the analytical methods were further improved, especially by the introduction of the nanospray technology (Wilm & Mann 1996), and also the interpretation of mass spectra became easier and more reliable with the aid of web-based databases (Mann & Wilm 1994). Mass spectrometry is, in addition, able to characterize post-translationally modified peptides (see below). Compared to Edman degradation, the determination of yet unknown sequences ('*de novo* sequencing') is much more difficult, since interpretation of primary experimental data depends heavily on database entries. The sensitivity of mass spectrometry as used for HLA ligand characterization is superior to any other analytical method that reveals structural details. Current technologies are able to successfully sequence peptides in the low femtomolar range on a routine basis, and sophisticated applications may even perform well in the attomolar range.

Alignment, motif determination and epitope prediction

All information obtained from pool sequencing is complemented by individual ligand characterization in order to establish a comprehensive peptide motif suitable for reliable epitope prediction. This process will be described in the following example. Table 1 shows 28 peptides presented by HLA-A*2402 molecules. The presence or absence of every amino acid in each sequence position is scrutinized and compared to information resulting from other sources, such as pool sequencing (Maier et al 1994), binding studies with synthetic peptides (Kondo et al 1995), listings of T cell epitopes, and the 3D structure of the MHC molecule. Most of the information provided by Table 1 is in agreement with previous findings: the anchor residues Y in position 2 (P2) and

TABLE 1 Natural HLA-A*2402 ligands used for establishing a prediction matrix

Sequence	*Protein*	*Position*
AYVHMVTHF	Testis-enhanced gene transcript	45–53
DYLKRFYLY	Matrilysin	37–45
EYPDRIMNTF	β tubulin	159–167
FYLEGGFSKF	Dual specificity phosphatase 6	130–139
FYPPKVELF	Multifunctional protein ADE2	121–129
GYGGGFGNF	Grancalcin	5–13
IYTKIMDLI	KIAA0877	24–32
KYISKPENL	FLJ12577	199–207
KYITQGQLLQF	Long chain fatty acid elongation enzyme	200–210
KYPDRVPVI	$GABA_A$ receptor associated protein	24–32
KYPENFFLL	NK cell activation protein	76–84
LYPQFMFHL	Sec23A	576–584
NYIDKVRFL	Vimentin	116–124
QYVPVIIHLI	Elongation of very long chain fatty acids	228–236
RYPDSHQLF	Ras–GAP SH3 binding protein	326–334
SYIEHIFEI	Phosphoprotein enriched in astrocytes 15	61–69
SYLPLAHMF	Long-chain-fatty-acid-CoA ligase 6	318–326
TYGEIFEKF	NADH-dehydrogenase subunit B14.5B	107–115
TYLEKAIKI	Ubiquitin C-terminal hydrolase 7	1092–1100
TYWVVYGVF	Polyposis locus protein 1	84–92
VYIEKNDKL	v-erb-b2 oncogene homologue 3	147–155
VYIKHPVSL	Proteasome subunit p31	131–139
VYISEHEHF	Cleft lip and palate associated tm protein	107–115
VYLKHPVSL	Proteasome 26S subunit non-ATPase 8	131–139
VYLPNINKI	KIAA0740	526–534
VYSHVIQKL	Serine dehydratase	277–285
YYIFIPSKF	Dead box protein	241–249
YYEEQHPEL	NK cell protein 4	107–115

L, F, I in P9, the preference for several amino acids in different positions (such as V in P1, P in P3, E in P4), and the general length of nine or, more rarely, 10 amino acids (Kubo et al 1994). Nevertheless, the set of 28 peptides seems not to be representative with respect to two features. First, a phenylalanine in P2 is not found in any of the ligands, but we know from pool sequencing, from binding studies, and from individual T cell epitopes that F plays a certain role in A*2402

ligands. Second, one ligand carries Y in P9, an amino acid that has not been found before to be important among A*2402-presented peptides. Thus, we have to notice that this number of ligands is still too small to represent all features of the peptide pool presented by a given HLA allotype, and we estimate that more than 50 sequences are required for comprehensive information.

After compiling all the information, the motif is translated into a computer-readable matrix as shown in Table 2. This matrix assigns definite values to every amino acid in each sequence position, but is not able to deal with mutual influences between amino acids within one sequence.

The capacity of the matrix is finally validated by two steps: first, the source proteins of natural ligands are screened, and experimentally determined sequences are expected among top-scoring peptides. These steps usually work with a success rate of more than 90% (data not shown), only the 11mer peptide in Table 1 escapes this kind of prediction, since the prediction patterns are strictly length-dependent and only available for 9mer and 10mer peptides. Second, HLA-A*2402-restricted T cell epitopes are compiled from the literature (e.g. as listed in the SYFPEITHI database), and their source proteins are subjected to epitope prediction. Table 3 lists the results of epitope prediction, including 50 HLA-A*2402-restricted T cell epitopes from 31 antigenic proteins of different origin. Such predictions can be reproduced using the SYFPEITHI epitope prediction programme, the A*2402 matrix has been incorporated into the latest update in September 2002. According to our definition, epitope prediction has been successful if a peptides ranks among the top-scoring 2% of peptides. For example, the sequence of a 500 amino acid protein harbours 492 possible nonamer peptides. Thus, the correct epitope should rank among the highest ten values. The far right column in Table 3 indicates that with the newly established epitope prediction, 38 epitopes would have been precisely predicted (76% reliability). If we only consider one epitope per antigen, epitope prediction from 28 out of 31 antigens was successful (90% reliability).

Modified peptides

Using routine procedures, it is hardly possible to characterize phosphorylated or glycosylated peptides. Their physicochemical properties impede their recovery from the HLA-bound peptide pool, since HPLC separation, Edman degradation and mass spectrometry are for different reasons not able to easily detect and analyse such modified peptides. Apart from reports on T cell recognition of glycosylated peptides (Ferris et al 1996) and qualitative description of glycopeptide presentation by HLA molecules (Kastrup et al 2000), we know a number of phosphorylated HLA ligands, identified again by Hunt and co-workers after employing special chromatographic procedures (Zarling et al 2000). More easily recognizable

TABLE 2 Motif pattern used for the prediction of HLA-A*2402-restricted T cell epitopes

(i) Nonamers

	Sequence position								
AA	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
A	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0
E	0	0	0	2	0	0	1	0	0
F	0	6	0	0	0	1	1	0	10
G	0	0	0	0	0	0	0	0	0
H	0	0	−1	0	1	1	0	1	0
I	0	0	2	0	1	1	0	0	10
K	1	0	−1	1	1	0	0	2	0
L	0	0	2	0	0	0	0	2	10
M	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0
P	0	0	2	1	0	1	0	0	0
Q	0	0	0	0	0	0	0	0	0
R	0	0	−1	0	1	0	0	0	0
S	0	0	0	0	0	0	0	0	0
T	1	0	0	0	0	0	0	0	0
V	2	0	0	0	1	1	0	0	0
W	0	0	0	0	0	0	0	0	0
X	0	0	0	0	0	0	0	0	0
Y	0	10	0	0	0	0	0	0	0

modifications are point mutations (Wölfel et al 1995) and frameshift mutations (Ronsin et al 1999), since they do not change the peptidic character of analytes. Other post-translational modifications have also been described (Skipper et al 1996, Pierce et al 1999), but it is impossible at present to exactly determine the overall quantity or ratio of modified peptides among all naturally presented HLA ligands. Interestingly, the presentation of phosphorylated peptides occurs more often by HLA-B*0702 molecules in comparison to HLA-A*0201molecules.

A question of quantity

The absolute copy number of a given MHC-peptide complex is important in several respects. From the analytical point of view, it decides whether a ligand might be

TABLE 2 (*Continued*)

(2) Decamers

	Sequence position									
AA	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
A	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0	0
E	0	0	0	2	0	0	0	1	0	0
F	0	6	0	0	0	0	1	1	0	10
G	0	0	0	0	0	0	0	0	0	0
H	0	0	−1	0	0	0	1	0	1	0
I	0	0	2	0	0	0	1	0	0	10
K	1	0	−1	1	0	0	0	0	2	0
L	0	0	2	0	0	0	0	0	2	10
M	0	0	0	0	0	0	0	0	0	0
N	0	0	0	0	0	0	0	0	0	0
P	0	0	2	1	0	0	1	0	0	0
Q	0	0	0	0	0	0	0	0	0	0
R	0	0	−1	0	0	0	0	0	0	0
S	0	0	0	0	0	0	0	0	0	0
T	1	0	0	0	0	0	0	0	0	0
V	2	0	0	0	0	0	1	0	0	0
W	0	0	0	0	0	0	0	0	0	0
X	0	0	0	0	0	0	0	0	0	0
Y	0	10	0	0	0	0	0	0	0	0

detectable or below the detection limit. From the cell biologist's viewpoint, it may indicate the turnover of the source protein within the cell, since proteasomal degradation represents a crucial step during antigen processing (Groettrup et al 2001). No matter if proteins are targeted to the proteasome by ubiquitylation after incorrect biosynthesis (defective ribosomal products: DRiPs, Yewdell et al 1996) or after having passed their lifetime in a functional state, the turnover rate seems to be more important for HLA-ligand creation than the overall amount of the respective source protein within the cell. From the medical point of view, the quality and the quantity of HLA-peptide presentation gives invaluable information about the state of a cell. Not only the presence or absence of HLA ligands derived from viral proteins tells about acute or latent infections but also tumour immunologists have great hope of using highly overpresented peptides from tumour antigens for

TABLE 3 Validation of the epitope prediction. Verification of the HLA-A*2402 motif pattern by prediction of all known HLA-A*2402-restricted CTL epitopes as listed in SYFPEITHI from their respective source proteins. Top 2%, the sequence is among the highest-scoring peptides in the respective protein. Parentheses indicate secondary epitopes from the respective protein

Source	*Sequence*	*Swissprot ID/ trembl Accession*	*score*	*rank*	*length*	*top 2%*
ART-1	EYCLKFTKL	O94864	24	1	414	+
ART4	DYPSLSATDI	Q9ULX3	22	3	412	+
ART4	AFLRHAAL	Q9ULX3	n.d.	n.d.	412	(−)
β catenin mut.	SYLDSGIHF	CTNB	24	1	781	+
CEA	TYACFVSNL	CEA5	22	5	702	+
CEA	QYSWFVNGTF	CEA5	20	12	702	(+)
CMV pp65	QYDPVAALF	PP65_	24	1	561	+
CMV pp65	VYALPLKML	PP65_	22	3	561	(+)
Cyclophilin B	KFHRVIKDF	CYPB	18	3	208	+
Cyclophilin B	DFMIQGGDF	CYPB	16	4	208	(+)
EBV EBNA3	RYSIFFDY	EBN3_EBV	n.d.	n.d.	812	−
EBV LMP-2	TYGPVFMCL	LMP2_EBV	24	1	497	+
EBV Rta	DYCNVLNKEF	BRL1_EBV	20	3	605	+
HBV core	EYLVSFGVW	CORA_HPBVA	22	1	211	+
HCV	AYSQQTRGL	POLG_HCVBK	22	11	3010	+
HER-2/neu	TYLPTNASL	ERB2	24	1	1255	+
HER-2/neu	RWGLLLALL	ERB2	12	n.d.	1255	(−)
HIV-1 (BRU) gag p17	KYKLKHIVW	GAG_HV1BR	12	n.d.	511	−
HIV-1 (BRU) gp41	RYLKDQQLL	ENV_HV1BR	25	1	963	+
HIV-1 (BRU) gp120	LFCASDAKAY	ENV_HV1BR	6	n.d.	861	−
MAGEA1	NYKHCFPEI	MAG1	21	2	309	+
MAGEA2	EYLQLVFGI	MAG2	24	1	314	+
MAGEA3	TFPDLESEF	MAG3	20	2	314	+
MAGEA3	IMPKAGLLI	MAG3	15	12	314	(-)
MDR p3	LYAWEPSFL	MRP3	21	10	1527	(+)
MDR p3	AYVPQQAWI	MRP3	21	10	1527	(+)
MDR p3	VYSDADIFL	MRP3	23	5	1527	+
Nicot. U08021	YYMIGEQKF	NNMT	22	2	264	+
p15	AYGLDFYIL	MA15	21	2	128	+
p53	EYLDDRNTF	P53	23	1	393	+
p53	TYSPALNKMF	P53	22	2	393	(+)
p53	NYMCNSSCM	P53	10	n.d.	393	(−)

TABLE 3 (*Continued*)

Source	*Sequence*	*Swissprot ID/ trembl Accession*	*score*	*rank*	*length*	*top 2%*
p53	AIYKQSQHM (?)	P53	2	n.d.	393	(−)
p53	TFRHSVVV	P53	n.d.	n.d.	393	(−)
PRAME	LYVDSLFFL	MAPE	22	2	509	+
Recoverin	AYAQHVFRSF	RECO	21	1	199	+
Recoverin	QFQSIYAKF	RECO	19	3	199	(+)
Recoverin	QFQSIYAKFF	RECO	16	9	199	(−)
SART1	EYRGFTQDF	O43290	19	4	800	+
SART2	DYSARWNEI	Q9UL01	21	8	958	+
SART2	AYDFLYNYL	Q9UL01	20	14	958	(+)
SART3	AYIDFEMKI	Q15020	25	1	963	+
SART3	VYDYNCHVDL	Q15020	23	4	963	(+)
Telomerase	VYAETKHFL	TERT	24	1	1132	+
Telomerase	VYGFVRACL	TERT	23	2	1132	(+)
Tyrosinase	AFLPWHRLF	TYRO	22	3	529	+
Tyrosinase	AFLPWHRLFL	TYRO	19	11	529	(−)
WT1	RWPSCQKKF	WT1	14	6	449	+
WT1	CMTWNQMNL	WT1	10	n.d.	449	(−)
Yo	AYRARALEL	PC17	20	6	443	+

diagnostic or therapeutic purposes. It is, however, a tedious process to determine the exact copy number of individual HLA ligands. Such a task can be performed by HPLC mass spectrometry with synthetic peptides used as calibrants. A different approach, used for example for the analysis of differences in peptide presentation between tumour cells and normal cells of the same tissue, determines the ratio of peptides presented by a pair of samples. Figure 1 shows the analysis of two HLA ligands compared between a colon carcinoma sample and the corresponding normal tissue from the same tumour patient. The peptide ESTGSIAKR is equally presented by tumour tissue and normal tissue, while the peptide DAAHPTNVQR is presented by HLA-A*6801 molecules in a significantly higher amount. The latter peptide is derived from β catenin, a protein which has been described as tumour-associated. Although the role of β catenin and its HLA-A*6801 ligand have not yet been elucidated, such differences in peptide presentation may in future contribute much to diagnostic or therapeutic strategies. Table 4 shows a number of HLA-A*6801-presented peptides and their presentation ratios as determined from the above-mentioned tumour.

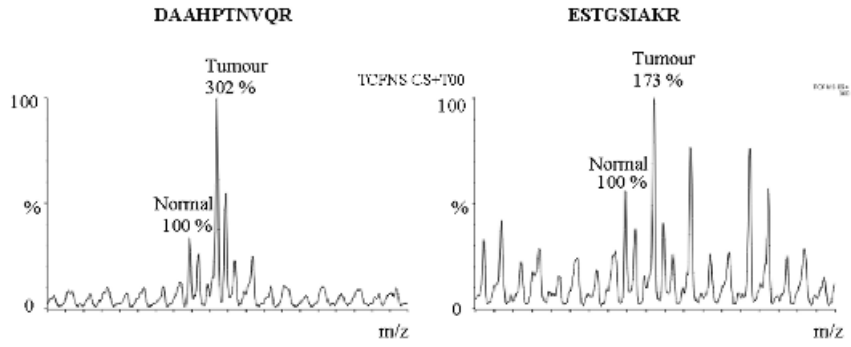


FIG. 1. Quantitative differences in HLA-A*6801-presented peptides between colon carcinoma and normal colon tissue from the same patient. From 7.2 g of normal tissue, 5.2 nmol of total HLA class I was immunoprecipitated (defined as 100%); from 7.0 g of tumour tissue, the total HLA class I yield was 9.8 nmol (188%). The peptide ESTGSIAKR was detected in an ratio tumour vs. normal of 173%, which corresponds to similar levels of presentation in both tissues. The peptide DAAHPTNVQR was detected at a higher level in the tumour.

Tissue-specificity and disease association

We have to face the fact that most HLA ligands we know have been extracted from quickly dividing, transformed cell lines with a high rate of metabolism. Therefore, it is not surprising that we know many HLA ligands from cell cycle-associated proteins, factors from signal transduction pathways, or proteins involved in protein biosynthesis. In contrast, only very few ligands have been characterized from normal, resting tissue with low division rates. Therefore, the majority of HLA ligands listed in the SYFPEITHI database might correspond to a transformed state in some way. Before we identify large numbers of MHC-presented peptides from normal tissue, we cannot judge the differences in the HLA-peptide repertoire between a normal state and a transformed state. It would also be interesting to know how big the differences between tissues of different origin are. Since B-lymphoblastoid cell lines (B-LCL) have commonly been used as a source of HLA-presented peptides, we know many HLA ligands that are specific for immune cells but that are not to be expected among HLA-presented peptides from solid tissues, and we cannot imagine with our present knowledge which part of the peptide pool is overlapping between different cell types. So, if we estimate the HLA peptide repertoire of one given cell at 10 000 different sequences, the HLA 'ligandome' of a human being might be much more complex and may contain far more than 100 000 unique peptides.

TABLE 4 HLA-A*6801 ligands from colon carcinoma and autologous normal colon tissue. An intensity value of 1.88 corresponds to equal presentation in tumour and normal tissue (see legend Fig. 1).

MW	*Sequence*	*Protein*	*Position*	*EMBL accession*	*Intensity tumour/normal*
1107.54	D A A H P T N V Q R	*β* catenin	115–124	X87838	3.02
948.50	E S T G S I A K R	Aldolase A	34–42	X05236	1.73
1003.51	D T A A Q I T Q R	MHC class I antigen (HLA-B)	136–144	U90245	1.61
981.50	E S G P S I V H R	Actin *β*	364–372	V00478	2.42
965.51	E A G P S I V H R	Actin α	366–374	D50029	1.51
1246.63	T A A D T A A Q I T R	MHC class I antigen (HLA-B)	133–144	U90245	2.19
1074.58	T T A E R E I V R	Actin alpha	204–212	D50029	1.45
884.54	A V A A V A A R R	Glucosidase II αsubunit	3–11	AF144074	1.91
898.55	V A V G V A R A R	Poly IG receptor	656–664	S62403	0.69
1025.58	D V S H T V V L R	Translocon-associated protein	88–96	X74104	1.07
1091.61	S I F D G R V V A K	Puatative membrane protein	88–97	AF274935	1.45
1075.18	D T I E I I T D R	Heterogenous nuclear RNP A2/B1	139–147	M29065	1.66
1030.58	E V T R I L D G K	SH3BGR3-like protein	23–31	AF304163	1.32
1048.61	T L G D I V F K R	Fatty acid-binding protein, liver	114–122	M10617	1.14

References

Barber LD, Percival L, Arnett KL, Gumperz JE, Chen L, Parham P 1997 Polymorphism in the alpha 1 helix of the HLA-B heavy chain can have an overriding influence on peptide-binding specificity. J Immunol 158:1660–1669

DiBrino M, Parker KC, Shiloach J et al 1993 Endogenous peptides bound to HLA-A3 possess a specific combination of anchor residues that permit identification of potential antigenic peptides. Proc Natl Acad Sci USA 90:1508–1512

Falk K, Rötzschke O, Stevanović S, Jung G, Rammensee HG 1991 Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature 351:290–296

Ferris RL, Buck C, Hammond SA et al 1996 Class I-restricted presentation of an HIV-1 gp41 epitope containing an N-linked glycosylation site. Implications for the mechanism of processing of viral envelope proteins. J Immunol 156:834–840

Garrett TP, Saper MA, Bjorkman PJ, Strominger JL, Wiley DC 1989 Specificity pockets for the side chains of peptide antigens in HLA-Aw68. Nature 342:692–696

Groettrup M, van den Broek M, Schwarz K et al 2001 Structural plasticity of the proteasome and its function in antigen processing. Crit Rev Immunol 21:339–358

Hunt DF, Henderson RA, Shabanowitz J et al 1992 Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. Science 255:1261–1263

Kastrup IB, Stevanović S, Arśequéll G et al 2000 Lectin purified human class I MHC-derived peptides: evidence for presentation of glycopeptides in vivo. Tissue Antigens 56:129–135

Kondo A, Sidney J, Southwood S et al 1995 Prominent roles of secondary anchor residues in peptide binding to HLA-A24 human class I molecules. J Immunol 155:4307–4312

Kubo RT, Sette A, Grey HM et al 1994 Definition of specific peptide motifs for four major HLA–A alleles. J Immunol 152:3913–3924

Maier R, Falk K, Rötzschke O et al 1994 Peptide motifs of HLA-A3, -A24, and -B7 molecules as determined by pool sequencing. Immunogenetics 40:306–308

Mann M, Wilm M 1994 Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem 66:4390–4399

Pierce RA, Field ED, den Haan JM et al 1999 Cutting edge: the HLA-A*0101-restricted HY minor histocompatibility antigen originates from DFFRY and contains a cysteinylated cysteine residue as identified by a novel mass spectrometric technique. J Immunol 163:6360–6364

Rammensee HG, Friede T, Stevanović S 1995 MHC ligands and peptide motifs: first listing. Immunogenetics 41:178–228

Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanović S 1999 SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50:213–219

Rötzschke O, Falk K, Deres K et al 1990 Isolation and analysis of naturally processed viral peptides as recognized by cytotoxic T cells. Nature 348:252–254

Ronsin C, Chung-Scott V, Poullion I, Aknouche N, Gaudin C, Triebel F 1999 A non-AUG-defined alternative open reading frame of the intestinal carboxyl esterase mRNA generates an epitope recognized by renal cell carcinoma-reactive tumor-infiltrating lymphocytes in situ. J Immunol 163:483–490

Skipper JC, Hendrickson RC, Gulden PH et al 1996 An HLA-A2-restricted tyrosinase antigen on melanoma cells results from posttranslational modification and suggests a novel pathway for processing of membrane proteins. J Exp Med 183:527–534

Van Bleek GM, Nathenson SG 1990 Isolation of an endogenously processed immunodominant viral peptide from the class I H-2Kb molecule. Nature 348:213–216

Wilm M, Mann M 1996 Analytical properties of the nanoelectrospray ion source. Anal Chem 68:1–8

Wölfel T, Hauer M, Schneider J et al 1995 A p16INK4a-insensitive CDK4 mutant targeted by cytolytic T lymphocytes in a human melanoma. Science 269:1281–1284

Yewdell JW, Anton LC, Bennink JR 1996 Defective ribosomal products (DRiPs): a major source of antigenic peptides for MHC class I molecules? J Immunol 157:1823–1826

Zarling AL, Ficarro SB, White FM, Shabanowitz J, Hunt DF, Engelhard VH 2000 Phosphorylated peptides are naturally processed and presented by major histocompatibility complex class I molecules *in vivo*. J Exp Med 192:1755–1762

Zwaveling S, Ferreira Mota SC, Nouta J et al 2002 Established human papillomavirus type 16-expressing tumors are effectively eradicated following vaccination with long peptides. J Immunol 169:350–358

DISCUSSION

Gulukota: You mentioned that some peptides were over-present and under-present. Is that the average of many experiments, or is that from a single sample?

Stevanović: The main problem is that we have just one experiment. This is always a very individual story. We have a tumour sample and because the amount of peptide is so low we cannot divide it. We are doing well if we can get peptides at all from these solid tumour samples. This was a preliminary list from just one experiment. We have been extracting solid tumours for seven years now, but this is the first time we attempted a quantitative comparison. Usually we end up with a list of 20 peptides and we have to screen through databases to find out which ones might be tumour associated. If we get a proto-oncogene, it is clearly tumour associated. But many other proteins are just over-expressed on the protein level and over-presented on the peptide level. This is what we try to establish, but it is always on an individual level. Perhaps we could compare different individuals if some peptides happen to be over-presented many times.

Perelson: Do you get your normal tissues from the same patients?

Stevanović: Yes. This is still a problem because we need rather homogeneous tissue, and this is difficult for many cell types.

Beck: Is there any programme in place to generate peptide sequences presented by tumour cells?

Stevanović: We still need more data. This won't happen until we get quantitative data, and this will take two or three more years. From our predictions we are not able to say anything about quantities of presented peptides.

De Groot: Could you say something about the clinical applications of your work?

Stevanović: At the moment we are still in the state of patient individual analysis. We would like to use these patient-specific peptides in therapy. We have to evaluate whether they are tumour associated or tumour specific, and if they are they could be used as a patient-specific vaccine. Then we would like to compose a cocktail of peptides that should address several MHC molecules and several antigens, because then we would have a good chance of avoiding tumour escape. We would like to vaccinate the patient with this cocktail, and the advantage is that we already know

that these peptides are presented by the autologous tumour. We haven't started these vaccinations yet, but they should take place in the next few weeks.

De Groot: Are you using whole antigen?

Stevanović: No, only peptides.

De Groot: This is in the context of the individualized vaccines, and the concerns about how feasible it is to do this. I am very interested in this because I would like to make DNA plasmids that contain the right set of MHC-restricted peptides or epitopes for certain individuals. It is the same idea, but it sounds like you are going patient specific rather than MHC specific. The other question is how much difference is there between patients? Do you see different antigens being presented? Is it worth taking colon cancer antigens from one patient and including that in a cocktail for everyone with A2?

Stevanović: Yes and no. We have analysed thoroughly the first two patients with renal cell carcinoma, and we saw that some of the HLA-A*0201-presented peptides were shared and the corresponding antigens were over-expressed in the respective tumours. Others are really specific for each patient. I wouldn't like to write a bill for what we are doing because this is very expensive at the moment. The analytical phase is expensive because we do analysis of HLA presented peptides (not too expensive) and gene expression profiling (more expensive). Therapy at the beginning is very expensive because we have to synthesize the peptides under GMP conditions. But if we find that some of the peptides are overlapping or shared between patients, this may make it possible to get most of the peptides from the shelf, which would cut costs.

De Groot: What about using tumour-infiltrating lymphocytes, like Steve Rosenberg is doing?

Rammensee: Generally, the peptides that Stefan finds are not recognized by T cells. We went to great lengths to find such T cells both in patients and healthy subjects. By and large we don't find T cells that recognize these peptides. If the patient's T cells would be effective against the tumour, the tumour would not be there.

Silva: Are different peptides being presented in different differentiation processes in the tumour? When the tumour starts growing the differentiation state of the cells is different. As it progresses, the phenotype of the cell changes.

Rammensee: Most likely, but this is very hard to follow.

Silva: If you want to vaccinate against a tumour, it will be difficult to know what its differentiation stage is.

Borras-Cuesta: I'd like to make a comment about predicting poor binders and good binders, because this is relevant to the field you are studying. Poor binders are important in cancer. It is important to be able to predict those, because it is from them that we will be able to induce an immune response against cancer. In some cases, such as HLA-A2, it has been described that if you put tyrosine in position 1,

then you greatly enhance binding (Tourdot et al 2000), although that doesn't come out of your original motif in the paper that you published in *Nature* (Falk et al 1991). We have tried this ourselves. We really can enhance the immunogenicity of these peptides just by putting a tyrosine in. The induction that you induce cross-reacts with a wild-type peptide, which is very good.

Rammensee: Not every clone will cross react. You will probably get a certain number of clones that will.

Borras-Cuesta: It recognizes the wild-type peptide presented by lysed cancer cells. We know this. We have replaced amino acid at position 1 by Tyr in many peptides, and it works. That is, the peptides become more immunogenic. The original paper did the same (Tourdot et al 2000). My question is, have you changed these in your prediction program now to see whether binding can be enhanced if tyrosine or phenylalanine are in position 1?

Stevanović: We do not look at binding data. Right from the beginning we decided only to take into account the natural ligands, so we don't care about the binding strength, and we don't care about binding studies with synthetic peptides. We just include the occurrence of amino acids in natural peptides. I have no idea what the score of tyrosine in position 1 in A2 is.

Borras-Cuesta: I have done it for you: I have reprocessed your data and shown that tyrosine is important in position 1.

Margalit: Is there a correlation between the gene expression and the peptide presentation?

Stevanović: From the gene expression analysis, from the many thousands of genes that are tested we usually get several hundred that are over-expressed more than fourfold. But the peptides we find, even if we find 77 like in the patient where we have been most successful so far, from those 77 ligands we have about 10 coming from the over-expressed genes. There is a long way to go from mRNA expression to peptide presentation. Hopefully we will find out in the next year or two whether there is some correlation.

Gulukota: I wanted to expand on what Hans-Georg Rammensee was saying about whether our natural antigenic peptides are among the top 5% of binders. If you are talking about therapy, you might want to look for under-represented peptides. The cancer already knows how to deal with the over-represented ones. The significance of whether a peptide binds strongly or weakly to an MHC complex depends on what it is we want to do with that. Natural ligands perhaps bind reasonably well. When we are looking at tinkering with the system in therapy, we can't just go with the flow of the biology: we probably need to look at other things. We might want to look for poor binders and see how we can exploit this.

Kellam: Many of these tumour antigens are expressed and function normally in cell developmental pathways. If you look at many stages of lymphomas and

leukaemias, the genes that are over-expressed reflect the stage of a normal B cell. If you are going to start to use peptide epitopes to target a tumour, whether it is solid, or a lymphoma or a leukaemia, are you not going to run the risk of an Elan-type scenario of actually removing a normal stage of cell development? This could result in acute or long-term toxicity.

Borras-Cuesta: They are monoclonal. If you have lymphoma, they are monoclonal, so you deal with that.

Kellam: But the vaccine is against a particular epitope from a normal host protein. It is not something exogenous.

Borras-Cuesta: In the case of myeloma you would have an immunoglobulin that will be exactly monoclonal. If you target this specifically, you will not target the others.

Kellam: If you target something that is not an immunoglobulin you would have a problem.

Borras-Cuesta: Most tumour antigens are self-antigens. That is why it is important to target the peptides that don't bind particularly well to the MHC because they have not been deleted during clonal selection. This is the point I was making before. It is important to address the question of which peptide antigens to target.

Rammensee: There are two different concerns here. One is that we would not get a T cell response, the other is that we would get a T cell response against an important host antigen expressed at a particular developmental stage. This latter point is a serious concern. We have tried to address it by looking at gene expression from all kinds of standardized tissue samples. On the other hand, those people who are using total tumour cell lysate would have the same concern. But usually the T cell responses are so weak that they do not attack the tumour, and there are usually no autoimmune complications in all sorts of vaccinations except with vitiligo upon immunization with melanocyte antigens.

Kellam: I quite agree. When you are looking in an acute model of whether the tumour regresses or not, this is not the same as looking over five years as to whether you get pathology associated with the long-term vaccination.

De Groot: Are you saying that you have done studies and you have seen some autoimmunity?

Rammensee: No, we have not done these studies, they are in the literature (Ludewig et al 2000).

De Groot: This would obviously be a problem if you are identifying self antigens and you are putting them back in with dendritic cells which are great at expressing them and with an adjuvant or cytokine.

Rammensee: This is what is done in many clinical studies, but just with one or two peptides.

Brusic: Tumour cells express sets of genes different to healthy cells. The same pattern has been observed in viral infection: the cell starts producing different products, not necessarily just viral proteins, and the immune systems starts

seeing these products. The amount of peptides shown on the cell surface is important for immune recognition. For self-antigens, the T cell clones are usually deleted. Researchers start by analysing proteins that induce immune responses. They start with longer fragments that induce immune recognition, and study them further to identify peptides of ideal length. There are many viral and cancer T cell epitopes described in the literature that don't conform to the proposed canonical motifs. They lack specific canonical amino acids at anchor positions. Eluted self-peptides usually have common anchor positions. This is a big problem for predictive modelling because our datasets are heavily pre-selected by the presence of anchor residues. These sets are used for assessing the quality of predictions resulting a self-justifying cycle. If I was a patient in need of a vaccine I would like to identify every peptide that can induce an immune response in my cancer, not only the few peptides that are best known. Our predictive models actually miss many true T-cell epitopes because of the pre-selection bias. Prilliman and Hildebrand from Oklahoma City extracted motifs using large quantities of HLA from bioreactors and actually found that for one molecule, B15, there are four different motifs. Can you comment on this? What should we do to obtain a more complete picture, rather than focusing on a subset of peptides that we can predict very well?

Rammensee: The aim should be to identify more than 10 000 peptides on one MHC molecule. They might fall into discernible groups that may identify motifs.

Brusic: There must be a way to identify MHC-binding peptides faster than using a rather pedestrian identification of a single peptide at a time. This is a major role of immunoinformatics.

Margalit: If they extract the peptides from the MHC molecules, these are the peptides that bind there. These are the facts.

Brusic: But are we extracting only the most abundant peptides, which are likely a consequence of the high quantity of protein being produced inside the cell?

Borras-Cuesta: The possible combinations are terrific. This is very difficult. I am convinced that all the prediction methods currently available predict only part of the picture. It is important to increase this number of peptides.

Rammensee: Perhaps the quantitative approach would help here. Darren Flower, would it be desirable to include in your database parameters such as the copy number of this peptide on a given cell to make it more complex with regard to information on the hierarchy of T cell responses?

Flower: If the data are out there, then they could be incorporated.

De Groot: Could you imagine collecting all those data?

Rammensee: We are talking about the optimum approach here.

Perelson: It would be fantastic if someone were to do this.

Lybrand: There is another option for expanding your repertoire. You have the QSAR parameters: have you thought about proposing other peptide

motifs that should be physically compatible on the basis of your QSAR profile?

Flower: Again, this is the idea of heteroclitic residues where you are trying to increase affinity. We haven't done this yet. It still comes back to the bias problem, that our models are still built with biased data; there is a lot of information missing. You could make some extrapolations and try to test them. The same is true with experimental design, trying to get away from the bias in the data sets that have been extracted from the literature. If you take a motif initially and generate a diverse set of peptides, you could test them and look for high binders. If you then look at the changes to the main anchors and run this process iteratively. This should allow you to get a much broader and deeper model.

Brusic: The strategy is to start in a grey area, do computational analysis, and follow this up with experimental validation.

Flower: You can't just do this once. This has to be run for several cycles in order to explore all the possibilities. What we have at the moment is a very biased data set, as it is self-reinforcing.

DeLisi: There is a more fundamental approach. This is to take each of the pockets, and on the basis of the amino acids in the pockets and the type of variability that exists, there is a finite number of families and superfamilies into which all MHCs can be put. On the basis of the pockets you can accommodate certain types of amino acid side chains, so you wind up with a set of families and superfamilies of MHC which accommodates a certain number of side chains in each pocket. You know what these are on the basis of a detailed physicochemical analysis of the interaction between particular side chains and particular pockets. This produces a very large combinatorial set, but we know exactly which amino acids could be accommodated by which pockets. In principle, you could predict the entire repertoire this way. We did this in a paper in 1998 (Zhang et al 1998) and we validated the results with the then available data: we had about 95% efficacy. I haven't pursued this approach, but there are no data biases in it.

Rammensee: You have the bias that your pockets are defined on high binding peptides.

DeLisi: The pockets are defined on the basis of crystal structures.

Rammensee: And the crystals are made of high binding peptides. There is no crystal structure of an MHC with a poorly binding peptide.

DeLisi: How much does the binding of the peptide bias which amino acids are in the pocket? It may bias the structure of the peptides a little bit.

Rammensee: Some reports have suggested that there are alternative binding frames for the low affinity peptides.

DeLisi: If you take the peptide out of the pocket you are left with an empty MHC.

Rammensee: The pocket might change.

DeLisi: We can take that into account. We take into account which side chains in the pocket are flexible. This could be done.

Lybrand: This is the strategy that we have been trying to use for eight or nine years now. I agree that this is a much more unbiased way to attempt to map out a potential repertoire of binding motifs. But there are two issues that have frustrated us to some extent. The first is that this kind of analysis is predicated on higher affinity binding of ligands to a target site than we are looking at here. There is a tendency to over emphasize the nature of the compatibility of the anchor residues with the pockets. We have had the best success in using this strategy in telling us what kinds of anchor residues are prohibited, but somewhat less success in telling us the full range of anchor residues that are OK. This is what we are focusing on here: we don't want the really good binders. We can tell you what is optimal and what is prohibited in these anchor pockets, but we cannot give you as good a feel for the intermediate kinds of anchors that would give the less good binders that people would like to explore.

DeLisi: I am not sure what the problem is for you. It depends on how good the free energy function is.

Lybrand: The other issue I have noticed more recently is that as we have begun to get a wider range of peptide–MHC crystal structures we are seeing a little more localized structural variation in these anchor pocket regions than I would have anticipated three or four years ago. A couple of these we have actually been able to predict successfully; others we never would have predicted successfully. We have seen side chains swing into different orientations and affect the local nature of the anchor pocket itself. I agree, though, that there is not a lot of structural variation here, so we are not looking at some highly variable target, which makes the philosophy a very attractive one. There is a little more variation than we would have hoped several years ago. It is not quite as simple an exercise as we anticipated, but it is still an appealing strategy to pursue.

Perelson: Stefan Stevanović, could we return to your estimate of 1000–10 000 peptides potentially being able to bind to a given MHC. Can you explain how you reached this number?

Stevanović: There may have been some misunderstanding. This is not the number of peptides that is able to bind to the MHC, but instead is the number actually presented by one cell. It has been calculated that millions of different peptides can bind to one certain kind of MHC.

Brusic: What is the difference in expression levels between a normal cell and an activated cell?

Stevanović: You would probably find a different pattern of MHC-presented peptides, but there are no data.

Perelson: The number presented by one cell is limited by the number of MHCs expressed per cell.

Stevanović: This is assuming that there are around 100 000 MHC copies per cell and they are just presenting between 1000 and 10 000 different epitopes.

Flower: Has this been measured?

Stevanović: It has been measured in several experiments. The highest number I have seen is 680 000 MHC molecules in one cell; the lowest is 50 000. We know that tumour cells have a tendency to reduce the number of MHC molecules expressed. For our peptide copy numbers there are also data from viral epitopes: some have 100 copies, others 500. For the peptide SYFPEITHI itself we know that it is 5000 copies.

Flower: So there are 1000–10 000 different peptides per cell. But you can only identify 77 of those, so how do you know there are many more?

Stevanović: Many papers from Donald Hunt's group have shown this (Hunt et al 1992, Luckey et al 2001). He pioneered mass spectrometry analysis of MHC ligands. From his profiles he identified peptide peaks but he couldn't sequence them because the intensity was too low. He estimated that there were 2000 of them. Other groups have made estimates that are in this range.

Gulukota: I'd like to get back to the computational discussion. We have discussed motifs, and whether *in silico* analysis could produce a list of peptides that could bind. If we look at the other side of this, are there any experimental strategies that we could adopt which would enrich for the low copy peptides? For example, if you had antibodies to the high copy peptides, could you pull them out so that what is left in the system is an enriched population of low copy peptides?

Brusic: There is a strategy involving bioreactors where sufficient yield of low quantity peptides can be produced. We have a formidable task — the complete number of peptides known to bind to HLA alleles is smaller than the diversity of peptides expressed by a single cell. We are still in the dark ages.

Littlejohn: This area looks like it is ripe for a functional genomics approach. You could take protein chips and wash the MHC peptide across them in a high-throughput analysis. I would synthesize every 9-mer, and bind them to a microarray. I would then wash MHCs across and see which ones bound. Is this a ridiculous idea?

DeLisi: No, we are developing assays to do exactly that. You need the right optical monitoring system.

Rammensee: There was a paper back in 1989 (Bouillot et al 1989) where many peptides were put on a plastic surface, and soluble MHC molecules were added. The result was the MHC molecules didn't have any reasonable peptide binding specificity! This was published in *Nature*.

Perelson: If one calculates the number of possible 9-mers it is 5×10^{11} peptides. What are the current estimates of how many of these a given MHC can bind?

Brusic: It is approximately 1%.

Perelson: How is this figure derived?

Petrovsky: By making overlapping peptides and then measuring binding.

Rammensee: Charles DeLisi, could you describe the chip you are working on?

DeLisi: Right now we are focusing on the engineering: how to fabricate arrays rapidly. We are doing both nucleic acids and peptides. We do *in situ* synthesis. The reason this isn't done much is because it is very expensive. If you want to do *in situ* synthesis you need physical masking at each step, and that becomes very expensive. But we do it in a different way. In a morning we can have a chip with 100 000 different pixels, with more than 1 000 000 oligos at each pixel We don't do random synthesis of the peptides because there has to be some sort of intelligent selection. The status is we can make the peptide arrays and we are characterizing them right now.

Rammensee: Isn't the problem that if you have the peptide on the surface, MHC has no access.

DeLisi: It has to be spaced. The same problem occurs with nucleic acids: we need a spacer of the right length. The peptides need to be far enough apart to avoid interference. This needs to be taken into account.

De Groot: Can you then add MHC and see what binds?

DeLisi: Yes, you can also do other things like running a phage display library over it. If you take a whole genome, for example, you could take a proteomic strategy seeing what the protein distribution is in a cell. If you have the whole genome you can select peptides that tend to be in surface proteins. You place these on the array and run a phage display library over it and then for those peptides that bind, you then have phage which binds and will therefore cross react with the native protein. Now you have the whole array of phage and you know which phage binds to which protein. You then plate the phage and use that as an assay for your protein distribution. The engineering bit turns out to be complicated because there are some purification steps that require precision engineering. But this is all technology that is becoming available in the next year or two.

Rammensee: From a purist's perspective, I would say that the way to solve this problem is to try to analyse all the 10 000 peptides on one cell. This would require us to improve our methodology and instrumentation. For many years we have been looking for the reverse translatase which makes RNA out of the peptide. Then we could amplify the RNA and easily identify a single peptide copy!

Kesmir: Stefan's data represent a wonderful test set for all these predictions we have been talking about. He also mentioned that just to get the number of copies of peptide per cell it will take two to three years. Could you explain a little more about why the quantitative data will take so long? It would be great to test any method on those data, because they are direct presentation data.

Stevanović: The main problem is just experimental details. You have to be sure that you get the peptides in a quantitative way, and this is very difficult. Usually if we try to get peptides in a quantitative way by mass spectrometry this is not

possible. In order to quantify the peptides you need to modify the peptides, and the modification step causes some to be lost.

Kesmir: I hope that there are also other experimental groups that will pick up this method.

Stevanović: It was easier in the early days when we did Edman degradation. This is a method used for quantitation so you can put all the peptides you get from the cell in the sequencer. But you can only sequence five or six peptides per cell by Edman degradation. Below the picogram range mass spectrometry is needed and this is not quantitative.

Rammensee: Coming back to the problem of predicting TCR recognition of the peptide, we know a couple of CDR sequences in Marie-Paule's database, and we know a few of them are from TCRs with defined specificity. Are there enough data so far to compare the databases of CDR3 sequences with that of the peptides recognized by the TCRs? If not, which direction should we take in this area? The aim would be to predetermine the specificity of a TCR, or to find a motif in the CDR3 sequence that would tell us which peptide is recognized.

Lefranc: The approach which consists in identifying specific T cells is more efficient. We don't have enough data to make predictions on CDR3. What we are doing now in IMGT is to try to put together the amino acid sequence from the CDR3, the peptide and from the MHC, so people can know exactly that an amino acid is coming from a CDR3 of such a length and in such an environment, and that it is in contact with such an amino acid of the peptide and MHC. Tools are developed which will allow queries based on the amino acid properties and polymorphisms and on the amino acid positions according to the IMGT unique numbering in the TR CDR3 and in the MHC.

References

Bouillot M, Choppin J, Cornille F et al 1989 Physical association between MHC class I molecules and immunogenic peptides. Nature 339:473–475

Falk K, Rötzschke O, Stevanović, Jung G, Rammensee HG 1991 Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules. Nature 351:290–296

Hunt DF, Henderson RA, Shabanowitz J et al 1992 Characterization of peptides bound to the class I MHC molecule HLA-A2.1 by mass spectrometry. Science 255:1261–1263

Luckey CJ, Marto JA, Partridge M et al 2001 Differences in the expression of human class I MHC alleles and their associated peptides in the presence of proteasome inhibitors. J Immunol 167:1212–1221

Ludewig B, Ochsenbein AF, Odermatt B, Paulin D, Hengartner H, Zinkernagel RM 2000 Immunotherapy with dendritic cells directed against tumor antigens shared with normal host cells results in severe autoimmune disease. J Exp Med 191:795–804

Tourdot S, Scardino A, Saloustrou E et al 2000 A general strategy to enhance immunogenicity of low-affinity HLA-A2.1-associated peptides: implication in the identification of cryptic tumor epitopes. Eur J Immunol 30:3411–3421

Zhang C, Anderson A, DeLisi C 1998 Structural principles that govern the peptide-binding motifs of class I MHC molecules. J Mol Biol 281:929–947

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

HLA nomenclature and the IMGT/HLA Sequence Database

Steven G. E. Marsh

Anthony Nolan Research Institute and Department of Haematology, Royal Free & University College Medical School, Hampstead, London NW3 2QG, UK

Abstract. Early in their study it was recognized that the genes encoding the HLA molecules were highly polymorphic and that there was a need for a systematic nomenclature. The result was the WHO Nomenclature Committee for Factors of the HLA System, which first met in 1968, and laid down the criteria for successive meetings. This committee meets regularly to discuss issues of nomenclature and has published 16 major reports documenting firstly the HLA antigens and more recently the genes and alleles. The standardization of HLA antigenic specificities has been controlled by the exchange of typing reagents and cells in the International Histocompatibility Workshops. Since 1989 when a large number of HLA allele sequences were first analysed and named, the job of curating and maintaining a database of sequences has been of prime importance. In 1998 the IMGT/HLA database became the official repository for HLA sequences. In addition to the nucleotide and protein sequences the database contains information of the cell from which the sequence was obtained. The database which provides tools for sequence analysis and the submission of new data, is updated quarterly and now contains over 1500 HLA allele sequences.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 165–176

Many of the advances in the HLA field have come about through the collaborative International Histocompatibility Workshops (IHWs). The first of these took place at Duke University, Durham, USA in 1964, and workshops have taken place every three to five years since with the most recent, the 13th, taking place in Victoria, Canada in 2002. In the early days the numbers of participants was small, only sixteen laboratories took part in the 1st Workshop, where they compared the typing of a panel of eight cells using seven different techniques. However, by the time of the 12th Workshop in 1996, over 400 laboratories world-wide were participating in a variety of different projects using many different molecular based techniques and typing thousands of samples.

Early in the study of HLA, the potential complexity of the system was beginning to be recognized and the need for standardized nomenclature understood. This was

felt during both the 1st IHW in 1964 and again at the 2nd IHW in 1965 at the University Hospital, Leiden, The Netherlands, where it became apparent that different groups were each using their own local designations to describe the same antigens. During the 2nd IHW a committee was formed to discuss nomenclature. It met only once believing that the time was not ripe to decide on a final nomenclature and suggesting that only provisional terms be used. The report of this meeting signalled to the community that the need for a standard nomenclature had been recognized, and was no more than a single sentence: '*The question of nomenclature of the leukocyte antigens has been raised during the workshop. An advice on this matter will be formulated by a committee on nomenclature, which has been formed during this Workshop*' (Bruning et al 1965). During the 3rd IHW in Torino, Italy, in June 1967 the issue of nomenclature was discussed again, and following a second meeting in Williamsburg (USA) in September, while still awaiting the formation of an official nomenclature committee, the main investigators in the field '*agreed to use the term HL-A for indicating the major system of leukocyte antigens (previous names: Du-1, Four, Hu-1, LA etc)*' (Amos 1968; Nomenclature Committee 1967). Contrary to popular belief the name assigned, 'HL-A' was not an abbreviation for 'Human Leucocyte Antigen' or 'Human Locus A' but simply as a contraction of the 'H' from 'Hu-1' system of Dausset and 'LA' from the system named by Payne and Bodmer (Amos 1999). In September 1968 under the auspices of the World Health Organisation (WHO) the first meeting of the 'WHO Leucocyte Nomenclature Committee' took place in New York. This meeting was recorded and a full verbatim account of the meeting was reported later (Walford 1990). The first eight serologically defined HL-A antigens were named at this time HL-A1 through HL-A8. These official names together with their previous locally assigned designations, as used in 10 different laboratories, were listed in the first full Nomenclature Report (WHO Nomenclature Committee 1968). The report also listed guidelines on the use of the new nomenclature and on the criteria used in the assignment of new antigens.

After the 4th IHW which took place in Los Angeles in 1970, a further four antigens were deemed worthy of an official designation HL-A10, -A11, -A12, -A13. For some reason HL-A9 is not listed at this time, however, the nomenclature report makes reference to its existence and that it was readily recognizable, suggesting that this antigen had been named in the intervening period between the workshops (WHO Nomenclature Committee 1970). Following the 5th IHW in Evian, France in 1972, the Nomenclature Committee announced that the definition of a histocompatibility antigen would pass through four stages. Firstly, a new specificity would be detected by a laboratory and given a local designation. Secondly, if this specificity were to be confirmed by several of the reference laboratories, it would be given a provisional number preceded by the prefix 'W'. In the third stage, when all the reference laboratories had reached a firm

agreement on the definition of the new specificity, an HL-A number would be assigned. In the fourth stage, a chemical or molecular analysis would allow the HL-A specificity to be confirmed (WHO Nomenclature Committee 1972). Ten new HL-A specificities were listed in this report, each of which bares the new 'W' prefix, indicating that the antigen had 'Workshop' status. It was also recognized by this time that some specificities, for example HL-A9, appeared to represent a cross-reactivity between two component antigens. The committee introduced the concept of a 'Broad' specificity, such as A9 and its components, later to be termed 'Splits', which were named AW23 and AW24.

It had been evident for some time that the relationship between different HL-A antigens was complex, and that the original serologically defined specificities of this system were being assigned to two separate series (first or LA, and second or FOUR) corresponding to two linked genes, each with multiple alleles. As such these two genes would require a separate nomenclature. Following the 6th IHW in Århus, Denmark in 1975, it was decided to remove the hyphen from the name HL-A, and use HLA as a designation of the system (WHO IUIS Terminology Committee 1975). This was followed by a hyphen used as a separator, before a gene designation A, B, C, D etc. Hence the antigens defined previously were assigned either to the HLA-A (previously LA or first) or HLA-B (previously FOUR or second) gene. The specificities 1, 2, 3, 9, 10 and 11 became HLA-A1, -A2, -A3, -A9, -A10 and -A11; specificities 5, 7, 8, 12, 13, etc. became HLA-B5, -B7, -B8, -B12, -B13. The use of a lower case 'w' to indicate a provisional specificity was retained with the 'w' being inserted between the gene name and antigen number, hence HLA-Aw23. Once antigens had been verified successfully it could be upgraded to full HLA status by omission of the 'w'. In addition at this time two new genes were recognized and named the HLA-C and HLA-D loci. A total of 51 different HLA antigens had been recognized and assigned official designations by 1975.

The HLA Nomenclature Report published after the 7th IHW in Oxford, UK in 1977, saw the introduction of the HLA-DR locus. The designation DR for 'D' related, indicated that these serological specificities were in some way related to the HLA-D specificities which had previously been defined using the cellular technique of Mixed Lymphocyte Culture (MLC) (WHO Nomenclature Committee 1978). The notation used to represent antigen splits was revisited at this time with the suggestion that the broad antigen name should follow the split name in parenthesis; for example Aw23(9), where Aw23 was a split of the A9 antigen. Although the numbers 4 and 6 had been held in reserve since 1968 for the 4a and 4b specificities, it was not until the 1977 report that these were officially named Bw4 and Bw6 and were recognized as public epitopes being present on all of the HLA-B antigens.

The 1980 Nomenclature Report included only a handful of new antigens and saw no major additions or changes to the HLA nomenclature; a total of 92 antigens were listed (WHO Nomenclature Committee 1980). The 1984 Nomenclature Report, published after the 9th IHW in München, Germany, saw the assignment of two new HLA genes, HLA-DQ and HLA-DP (WHO Nomenclature Committee 1985). The newly assigned DQ specificities, DQw1, DQw2 and DQw3 were defined by serological techniques; the six new DP antigens, DPw1 to DPw6 were defined using the cellular assay Primed Lymphocyte Typing (PLT). In addition two new DR antigens were named DRw52 and DRw53. At the time it was unclear whether these represented public epitopes on the DR molecule in an analogous way to the Bw4 and Bw6 epitopes of HLA-B. It was later shown that these were the products of secondary HLA-DR genes. By this time an elementary map of the HLA region had been established and the first HLA genes had been cloned, and it was clearly understood that the HLA class II molecules consisted of two polypeptide chains whose genes were both located within the HLA region. The suggestion was made that the genes for the separate chains be called DRA and DRB etc.

In 1987 following the 10th IHW in Princeton, USA a molecular nomenclature for both genes and DNA allele sequences was introduced with the recognition that many pseudogenes were also located in the HLA region (Bodmer et al 1989). Expanding the previously suggested notation, the HLA genes were given official names, the gene encoding the DRα chain was called DRA. Several different genes encoding DRβ chains had been identified and were named HLA-DRB1, -DRB2, -DRB3 and -DRB4. The HLA-DRB2 gene was shown to be a pseudogene. In addition to naming many new genes, it was recognized that many of the antigen specificities previously defined by serology, such as HLA-A2, could be subdivided still further by DNA sequencing. Four different A2 sequences were named at this time, A*0201, A*0202, A*0203 and A*0204. The asterisk was used as a separator between the gene name, and the four-digit code used to distinguish between the alleles. The first two digits indicating the HLA antigen encoded by the allele and the second two digits indicating that number of the allele in that series, where each allele differs from the others by at least one nucleotide substitution that changes the amino acid sequence of the encoded protein. A total of 12 HLA class I alleles and nine class II alleles were named at this time.

The meeting of the HLA Nomenclature Committee in 1989 was the first to take place between workshops and recognized the need to assign official names to the many new HLA allele sequences that were being published (Bodmer et al 1990). A total of 56 class I and 78 class II alleles were named in the report of this meeting. It had become necessary to emphasise the need to deposit the newly discovered sequences in an appropriate database, and that this would need to be continually updated.

Since 1989 the HLA Nomenclature Committee has continued to meet every one to two years to establish further guidelines for the naming of HLA genes and alleles (Bodmer et al 1991, 1992, 1994, 1995, 1997, 1999, Marsh et al 2001, 2002). The HLA allele names were first extended to five digits in 1990 to allow for the discrimination of alleles differing only in non-coding (synonymous) substitutions within the coding sequence (Bodmer et al 1991). In 1995 they were again extended to seven digits to allow for the naming of alleles which differed only in introns or the 3′ or 5′ regions of the gene (Bodmer et al 1995). Then in 2002, due to the increasing number of alleles being described, an additional digit was inserted between the four and fifth digits to allow for more than nine alleles differing only by synonymous substitutions (Marsh et al 2002). Aside from extending the number of digits used to code for the different alleles, and the adding of optional suffixes to indicate whether an allele is null (an N), lowly expressed (an L) or only translated in a soluble form (an S), the nomenclature used for alleles has changed little since it was first introduced in 1987. The greatest development has been the dramatic increase in the numbers of HLA allele sequences discovered in this time. By 2002 the number of alleles that had been assigned had steadily grown to over 880 HLA class I alleles and over 600 class II alleles (see Fig. 1).

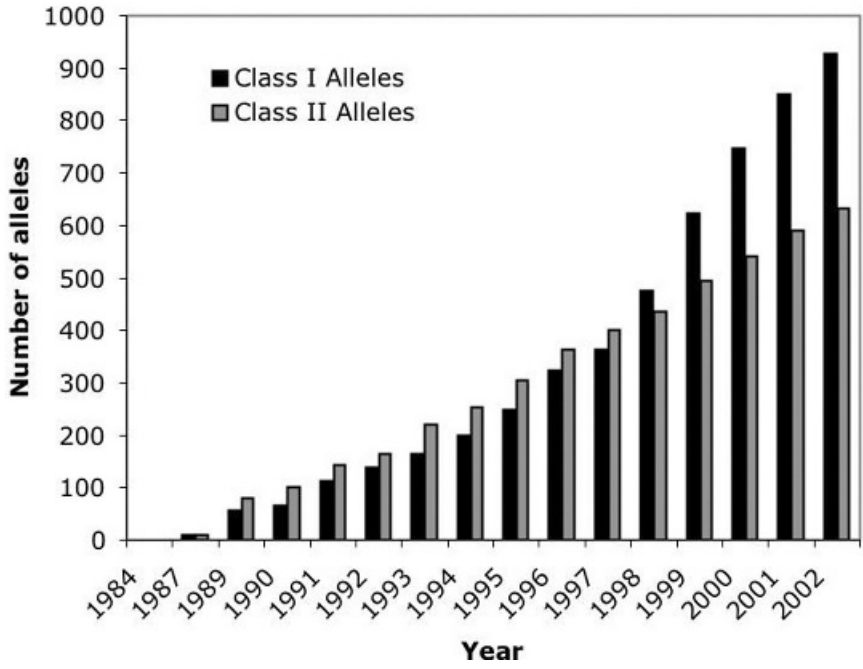


FIG. 1. The number of HLA class I and II alleles officially recognized since 1984.

It became apparent as early as 1989 that the analysis and assigning of official names to alleles could not wait for periodic histocompatibility workshops or even annual Nomenclature Committee meetings, and so began the process of daily assessing newly defined HLA allele sequences. This work was carried out by Julia Bodmer and Steven Marsh at the Imperial Cancer Research Fund (ICRF) in collaboration with Peter Parham at Stanford University. It was out of the need to record and manage the HLA sequence data being submitted to the Nomenclature Committee that the first incarnation of an HLA Sequence Databank (HLA-DB) emerged (Marsh & Bodmer 1993).

Periodically HLA class I (Arnett & Parham 1995, Mason & Parham 1998, Zemmour & Parham 1991, 1992) and class II (Marsh 1998, Marsh & Bodmer 1990, 1991, 1992, 1994, 1995) sequence alignments were published in a variety of journals and by 1995 the numbers of new alleles being reported warranted the publication of monthly nomenclature updates (Marsh 1995), something which continues to this day. Also by 1995, the expansion of the Internet and the introduction of the World Wide Web (WWW) saw the first distribution of the HLA sequence alignments from the web pages of the Tissue Antigen Laboratory at the ICRF. This work transferred to the Anthony Nolan Research Institute (ANRI) in 1996 where it continues today. In an effort to make the data held in the database available in a more accessible and interactive format the IMGT/HLA Database project was begun in 1997 as part of a European collaboration involving the ICRF, ANRI and the European Bioinformatics Institute (EBI) who maintain the European Molecular Biology Laboratory's nucleotide sequence database (EMBL) (Robinson et al 2000, 2001). The work was initially funded by grants from the European Union, BIOMED1 (BIOCT930038) and BIOTECH2 (BIO4CT960037), awarded to the ICRF as part of the International ImMunoGeneTics (IMGT) databases project (Giudicelli et al 1997, Lefranc 2001, Ruiz et al 2000). The IMGT database project contains a number of distinct databases specializing in sequences of immunological interest. The IMGT/HLA database was first released in 1998, the database combines the sequence data and information previously provided to the WHO Nomenclature Committee for Factors of the HLA System and the additional data found in the original EMBL/GenBank/DDBJ entries. The database can be accessed from *www.ebi.ac.uk/imgt/hla*. The current release of the database, version 2.1.0 contains over 1580 HLA alleles and details of over 2700 cells that have been sequenced for one or more HLA alleles. The database is updated every three months and provides a suite of tools for analysing the nucleotide and protein sequences. Since 2000 the IMGT/HLA Sequence Database has been supported by the generous donations of a number of commercial companies, immunogenetic organizations and bone marrow registries.

Under a new initiative the Immuno-Polymorphism Database (IPD) was launched in 2003, as part of a collaboration between the ANRI and the EBI. The first two component databases of this project are the IPD-MHC project, a database of MHC sequences from a variety of different animal species, and the IPD-KIR database, a sequence database of the Human Killer-cell Immunoglobulin-like Receptors (KIR). These databases may be accessed from *www.ebi.ac.uk/ipd*.

The value of a database of HLA allele sequences to the user communities in transplantation, research and clinical practice, is critically dependant on the quality and accuracy of the information it contains. Even single nucleotide errors in transcribing and reporting sequences cannot be tolerated if the data are to be relied on. The job of maintaining and curating the database is thus of vital importance and necessarily requires meticulous attention to detail. However, the IMGT/HLA Sequence Database now goes well beyond just providing a list of sequences and provides a whole range of linked information such as details of the cell line from which the sequence was derived. The online database incorporated many tools for data retrieval, analysis and submission of new data and is regularly updated. With the completion of the Human Genome Project and the identification of many new polymorphic genes, the IMGT/HLA Sequence Database and the HLA Nomenclature, which has evolved over the past thirty years, is clearly a model of how this new polymorphic data can be managed.

Acknowledgements

I would like to thank James Robinson and Matthew Waller for their input and hard work on the daily maintenance on the IMGT/HLA Sequence Database. Peter Stoehr and colleagues at the EBI for their continued support of the database. I would like to acknowledge the support of the following organisations for the IMGT/HLA Database: Dynal, Biotest, Orchid Biosciences, the American Society for Histocompatibility and Immunogenetics (ASHI), the Anthony Nolan Trust (ANT), Celera Diagnostics, the European Federation of Immunogenetics (EFI), Forensic Analytical, Genovision-Olerup SSP, The Marrow Foundation, the National Marrow Donor Program (NMDP), One Lambda, and Pel-Freez Clinical Systems.

Appendix: access and contact

IMGT/HLA Homepage:	*http://www.ebi.ac.uk/imgt/hla/*
IMGT/HLA Submissions:	*http://www.ebi.ac.uk/imgt/hla/subs/submit.html*
Contact:	*hladb@ebi.ac.uk*
IPD-KIR Homepage:	*http://www.ebi.ac.uk/ipd/kir/*
IPD-MHC Homepage:	*http://www.ebi.ac.uk/ipd/mhc/*
Non-human primates:	*http://www.ebi.ac.uk/ipd/mhc/nhp*
Canines:	*http://www.ebi.ac.uk/ipd/mhc/dla*
Felines:	*http://www.ebi.ac.uk/ipd/mhc/fla*
Contact:	*ipd@ebi.ac.uk*

References

Amos DB 1968 Human histocompatibility locus HL-A. Science 159:659–660
Amos DB 1999 Why "HLA". In: Hahn AB, Rodey GE (eds) ASHI: the first 25 years (1974–1999). American Society for Histocompatibility and Immunogenetics, p 64–66
Arnett KL, Parham P 1995 HLA class I nucleotide sequences, 1995. Tissue Antigens 46:217–257
Bodmer JG, Marsh SGE, Albert E et al 1995 Nomenclature for factors of the HLA system, 1995. Tissue Antigens 46:1–18
Bodmer JG, Marsh SGE, Albert E et al 1991 Nomenclature for factors of the HLA system, 1990. Tissue Antigens 37:97–104
Bodmer JG, Marsh SGE, Albert E et al 1992 Nomenclature for factors of the HLA system, 1991. In: Tsuji T, Aizawa M, Sasazuki T (eds) HLA 1991. Oxford University Press, Oxford, p 17–31
Bodmer JG, Marsh SGE, Albert E et al 1994 Nomenclature for factors of the HLA system, 1994. Tissue Antigens 44:1–18
Bodmer JG, Marsh SGE, Albert ED et al 1997 Nomenclature for factors of the HLA system, 1996. Tissue Antigens 49:297–321
Bodmer JG, Marsh SGE, Albert ED et al 1999 Nomenclature for factors of the HLA system, 1998. Tissue Antigens 53:407–446
Bodmer JG, Marsh SGE, Parham P et al 1990 Nomenclature for factors of the HLA system, 1989. Tissue Antigens 35:1–8
Bodmer WF, Albert E, Bodmer JG et al 1989 Nomenclature for factors of the HLA system, 1987. In: Dupont B (ed) Immunobiology of HLA. Springer-Verlag, New York, p 72–79
Bruning JW, van Leeuwen A, van Rood J J 1965 Leucocyte Antigens. In: Amos B, van Rood J J (eds) Histocompatibility testing 1965. Munksgaard, Copenhagen, Denmark, p 275–283
Giudicelli V, Chaume D, Bodmer J et al 1997 IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 25:206–211
Lefranc MP 2001 IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 29:207–209
Marsh SGE 1995 Nomenclature for factors of the HLA system, update January 1995. Tissue Antigens 45:220–222
Marsh SGE 1998 HLA class II region sequences, 1998. Tissue Antigens 51:467–507
Marsh SGE, Albert ED, Bodmer WF et al 2002 Nomenclature for factors of the HLA system, 2002. Tissue Antigens 60:407–464
Marsh SGE, Bodmer JG 1990 HLA-DRB nucleotide sequences, 1990. Immunogenetics 31: 141–144
Marsh SGE, Bodmer JG 1991 HLA class II nucleotide sequences, 1991. Tissue Antigens 37: 181–189
Marsh SGE, Bodmer JG 1992 HLA class II nucleotide sequences, 1992. Tissue Antigens 40: 229–243
Marsh SGE, Bodmer JG 1993 HLA Class II Sequence Databank. Human Immunol 36:44
Marsh SGE, Bodmer JG 1994 HLA class II region nucleotide sequences, 1994. Eur J Immunogenet 21:519–551
Marsh SGE, Bodmer JG 1995 HLA class II region nucleotide sequences, 1995. Tissue Antigens 46:258–280
Marsh SGE, Bodmer JG, Albert ED et al 2001 Nomenclature for factors of the HLA system, 2000. Tissue Antigens 57:236–283
Mason PM, Parham P 1998 HLA class I region sequences, 1998. Tissue Antigens 51:417–466
Nomenclature Committee 1967 Nomenclature: HL-A. In: ES Curtoni, PL Mattiuz, RM Tosi (eds) Histocompatibility testing 1967. Munksgaard, Copenhagen, Denmark, p 449

Robinson J, Malik A, Parham P, Bodmer JG, Marsh SGE 2000 IMGT/HLA database — a sequence database for the human major histocompatibility complex. Tissue Antigens 55:280–287

Robinson J, Waller MJ, Parham P, Bodmer JG, Marsh SGE 2001 IMGT/HLA database — a sequence database for the human major histocompatibility complex. Nucleic Acids Res 29:210–213

Ruiz M, Giudicelli V, Ginestoux C et al 2000 IMGT, the international ImMunoGeneTics database. Nucleic Acids Res 28:219–221

Walford RL 1990 First meeting WHO Leukocyte Nomenclature Committee, New York, September 1968. In: Terasaki PI (ed) History of HLA: ten recollections. UCLA Tissue Typing Laboratory, Los Angeles, p 121–149

WHO IUIS Terminology Committee 1975 Nomenclature for factors of the HLA system. Bull World Health Organ 52:261

WHO Nomenclature Committee 1968 Nomenclature for factors of the HL-A system. Bull World Health Organ 39:483–486

WHO Nomenclature Committee 1970 WHO Terminology Report. In: Terasaki PI (ed) Histocompatibility Testing, 1970. Munksgaard, Copenhagen, p 49

WHO Nomenclature Committee 1972 WHO Terminology report. Bull World Health Organ 47:659–662

WHO Nomenclature Committee 1978 Nomenclature for factors of the HLA system, 1977. Bull World Health Organ 56:461–465

WHO Nomenclature Committee 1980 Nomenclature for factors of the HLA system, 1980. In: Terasaki PI (ed) Histocompatibility testing, 1980. UCLA Tissue Typing Laboratory, Los Angeles, p 18–20

WHO Nomenclature Committee 1985 Nomenclature for factors of the HLA system 1984. In: ED Albert, MP Baur, WR Mayr (eds) Histocompatibility testing, 1984. Springer-Verlag, Berlin, p 4–8

Zemmour J, Parham P 1991 HLA class I nucleotide sequences, 1991. Tissue Antigens 37: 174–180

Zemmour J, Parham P 1992 HLA class I nucleotide sequences, 1992. Tissue Antigens 40: 221–228

DISCUSSION

Beck: Are the data on these KIR genes available yet?

Marsh: Not yet. The MHC data for some species are on the web but not the KIR data. We are working on a KIR nomenclature report, and when this is finalized we will make the first set of alignments available.

Beck: Will this be a separate database?

Marsh: I am not sure yet. We will provide links to it from where we are at the moment. [Since these discussions took place, the IPD/KIR Sequence Database has been released and is available from the *www.ebi.ac.uk/ipd/kir*.]

Rammensee: You didn't mention the mouse. Is anyone working on the mouse?

Marsh: We currently have no plans to tackle the mouse MHC, although we are in contact with several other groups for the inclusion of MHC sequences from this species.

Lefranc: I tried for many years to get people involved with the mouse MHC, without success. Since there is such a need in that field our lab decided finally to get involved in the way we could. We implemented the cards for the mouse MHC 3D structures in IMGT/3Dstructure-DB (*http://imgt.cines.fr*). We also added tables on 'Correspondence between the mouse MHC nomenclatures' in IMGT Repertoire, available from the IMGT Home page (*http://imgt.cines.fr*).

Marsh: If we can find someone who is willing to curate mouse MHC data, then we will certainly work with them to provide a database structure and the tools to manipulate their data.

Brusic: We also have data on BoLA and SLA data. There was a problem with nomenclature. The same naming problem that appeared 30 years ago in the HLA field is being faced in BoLA and SLA. There is serious resistance towards standardizing names of MHC molecules for other organisms. The nomenclature issues need to be resolved first.

Marsh: We are working on the model system where we are in contact with nomenclature committees, with people who want to curate their own data and who have the agreement of the people working in the field. They each have their own nomenclatures which they are going to maintain and which all of them will use. This has worked for HLA because it goes back nearly 40 years. It has not always worked for the other species, but the structured database approach that we are planning should aid in the uptake and use of these nomenclatures.

Schönbach: For example, the BoLA (bovine lymphocyte antigens) nomenclature committee seems to promote their own existing nomenclature although it could be improved in terms of creating and maintaining a searchable database and compatibility with existing MHC nomenclature rules.

Marsh: The data we are curating and making available in the database are nucleotide and amino acid sequences, and as such we need a robust genetic nomenclature that meets the needs of such data, rather than relying on old serological definitions.

Rammensee: With the mouse, the Jackson Laboratory has some gene bank availability, but not of the structure of MHC molecules and their peptide specificity.

Petrovsky: I am sure they would be interested. I don't think anyone has approached them. Ed Leiter at Jackson Laboratory is generally very enthusiastic about any programme for classifying mice.

Lefranc: We need to always clearly indicate the strain for mouse. We have, in IMGT, indicated the strains for the immunoglobulin and T cell receptor genes but this was quite a heavy task. Authors often forget to indicate the strain when submitting their sequences to EMBL or GenBank so we had to go back to the literature. The list of the mouse strains quoted in IMGT Repertoire, with links to

Mouse Genome Informatics (MGI), is available from the IMGT Index (*http://imgt.cines.fr*) > Strains.

Kesmir: You don't have any information about the frequency of HLA alleles.

Marsh: That's true.

Kesmir: This would be useful information. Is it possible to get it somewhere else?

Marsh: We haven't linked to too many other databases yet because we are unsure of the quality of some of the other databases out there. I would like to include more links because this would improve the functionality of what we have. A new database is now beginning called 'Allele Frequencies Database' (*www.allelefrequencies.net*). This is a new database that is collecting HLA frequency data as well as other genetic markers such as KIR alleles and cytokine polymorphism frequencies. At the present time there are few data. They currently link to the IMGT/HLA database but we don't have reciprocal links back yet. This is partly because of the structure of their database, which requires a password for access. The bottom line is that for the 1500 or so alleles that we have there aren't well controlled good quality frequency data in a variety of different populations, which is what we would love to see. It would be nice to click on a button and get the A*0259 allele frequencies in every population around the world.

Kesmir: What about haplotype information?

Marsh: The haplotype information comes after the frequency data. The best data set is still from the 1991 HLA workshop, and this was based on serology for class I and some DNA typing for class II. They looked at over 100 populations. Nothing has happened globally since then to give us a similar sort of data set. The Allele Frequencies database is storing haplotype frequency data in addition to allele frequencies.

Littlejohn: What are your submission processes? Do people have to submit data to EBI/EMBL first and then you?

Marsh: Yes. The criteria were drawn up a long time ago when we weren't sure we would ever get funding for our database and how big it would grow. We never wanted to have the data only in this database and then find that it can't be curated. We therefore stipulated that people should submit their data to Genbank/EMBL/DDBJ as well as to us.

Littlejohn: Theoretically, once the MHC multispecies starts filling up it will also have phylogenetic analysis tools.

Rammensee: Is there any connection with the bone marrow transplantation repositories? So far 8 million people have been typed. Is there exchange of data?

Marsh: We supply sequence alignments to the major registries every quarter, and they support us, but the level of resolution in the registries is low or medium. Even if you were to try to generate frequencies for some of these alleles it is impossible from this data set. On the clinical side, with regard to matching between donors

and patients, there are big studies going on where we are looking at the level of matching that is being obtained between patients and donors in retrospective analysis, and following it up with clinical data.

Rammensee: I was thinking the other way round: it might be interesting to see the plain frequency of one allele or family in the bone marrow transplantation registry to get some idea of the frequency of an allele in the population.

Marsh: As most of the data in the registry are low or medium level, we don't achieve anything more than what was achieved in the 1991 workshop where everything was done at the serological level.

Margalit: I wanted to ask about the annotation of the non-synonymous substitutions. Does this have any practical implications, or is it just for phylogenetic analysis?

Marsh: In transplantation, because there is such strong linkage disequilibrium it can tell us what other genes we expect to be on the same haplotype. This is one issue. The second issue is that when you are designing reagents that are specific for single nucleotide changes we need to be able to identify and name those sequences so we can use, in some cases, two or three probes to a specific site because we have these known substitutions.

Beck: Synonymous substitutions affecting CpG positions might also be important if you consider methylation patterns.

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

From immunogenetics to immunomics: functional prospecting of genes and transcripts

Christian Schönbach

Biomedical Knowledge Discovery Team, Bioinformatics Group, RIKEN Genomic Sciences Center (GSC), 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan

Abstract. Human and mouse genome and transcriptome projects have expanded the field of 'immunogenetics' beyond the traditional study of the genetics and evolution of MHC, TCR and Ig loci into the new interdisciplinary area of 'immunomics'. Immunomics is the study of the molecular functions associated with all immune-related coding and non-coding mRNA transcripts. To unravel the function, regulation and diversity of the immunome requires that we identify and correctly categorize all immune-related transcripts. The importance of intercalated genes, antisense transcripts and non-coding RNAs and their potential role in regulation of immune development and function are only just starting to be appreciated. To better understand immune function and regulation, transcriptome projects (e.g. Functional Annotation of the Mouse, FANTOM), that focus on sequencing full-length transcripts from multiple tissue sources, ideally should include specific immune cells (e.g. T cell, B cells, macrophages, dendritic cells) at various states of development, in activated and unactivated states and in different disease contexts. Progress in deciphering immune regulatory networks will require the cooperative efforts of immunologists, immunogeneticists, molecular biologists and bioinformaticians. Although primary sequence analysis remains useful for annotation of new transcripts it is less useful for identifying novel functions of known transcripts in a new context (protein interaction network or pathway). The most efficient approach to mine useful information from the vast a priori knowledge contained in biological databases and the scientific literature, is to use a combination of computational and expert-driven knowledge discovery strategies. This paper will illustrate the challenges posed in attempts to functionally infer transcriptional regulation and interaction of immune-related genes from text and sequence-based data sources.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 177–192

The discovery of the DNA-double helix 50 years ago (Watson & Crick 1953) triggered a new wave of biology that for the first time enabled biologists to link biological phenomena to the function of particular genes. The main technologies that have driven this revolution are cloning, sequencing and the identification of

specific functions at the sequence, transcript, protein, cellular and whole organism level. The release of the human draft genome (Lander et al 2001, Venter et al 2001) and mouse genome (Waterston et al 2002) in combination with annotation efforts (e.g. HAVANA) and haplotype structure and linkage disequilibrium analyses (Daly et al 2001, Rioux et al 2001) have set the stage for analysing the genome using the transcriptome and vice versa. Extensively curated cDNA clone data (Kawai et al 2001; FANTOM2; Human-Invitational cDNA Annotation Jamboree), gene profiling (Shaffer et al 2001) and oligo arrays (Shoemaker et al 2001) facilitate deciphering of the genetics and multi-dimensional structure of molecular processes that regulate immune function.

What can the transcriptome tell us about the immunome?

The functional annotation of 21 076 mouse full-length cDNA sequences as part of the RIKEN mouse gene encyclopaedia project (FANTOM1, Kawai et al 2001) resulted in the identification of 15 295 genes of which 58.3% did not correspond to known genes. The value of the FANTOM cDNA collection lies in the fact that it includes predominantly full length transcripts, from the diversity of tissue sources used to construct the libraries, with an emphasis on enrichment with novel transcripts, multiplicity of splice variants and its expert human curation. Interestingly, 95 clones were confirmed to be in reverse orientation and these may present candidate antisense transcripts that are involved in gene regulation.

A comparison of *ab initio* predicted exons with novel FANTOM transcripts showed that only 21% of transcripts perfectly overlapped with GenScan predicted exons, 38% showed partial matches and 41% were not predicted. The disparity between the large transcriptional capacity of the mouse genome and number of algorithm-predicted exons was also confirmed by oligonucleotide arrays (Kapranov et al 2002) of chromosome 21 and 22, which contained about 770 predicted and characterized genes whereas the actual transcriptional capacity of these chromosomes exceeded the number of predicted genes by an order of magnitude. Therefore, gene structure and the transcriptional capacity of the genome cannot be fully captured without transcriptional data derived from physical clones. Moreover, these data question the capacity to predict regulatory elements such as promoters without transcriptional supporting data.

Defining immunomics

Since publication of the FANTOM1 analysis results 60 000 cDNA sequences have undergone computational analyses and human curation (FANTOM2; Okazaki et al 2002) while this manuscript was being prepared. With the release of the RIKEN mouse cDNAs and the complete mouse genome sequence, we now

have an extraordinary reference resource for computational and experimental studies of immune system transcriptional regulation, protein interactions, signalling and metabolism. The size of the transcriptome depends on the definition of transcription unit, tissue sources of cDNA libraries, and representation of mRNA transcripts. A transcriptional unit requires that a promoter determines the direction and start-site of a transcription unit and a termination site the end. Transcripts that arise by recombination events or rearrangement such as *Ig* and *Tcr* would therefore be excluded. If we define 'immunomics' as the molecular functions associated with all immune-related coding and non-coding mRNA transcripts, the number of immune-related transcripts is estimated to be several orders of magnitude higher than the number genes encoding them. If we extrapolate the estimate of Arnone & Davison (1997) that each gene interacts with four to eight other genes and is involved in up to 10 different biological functions it is clear that functional transcript-based diversity and complexity is simply enormous. This poses problems for bioinformatics analyses and modelling of transcriptional networks. Before I highlight the biological effects consequent upon transcriptional diversity, some caveats relating to the current data need to addressed.

Caveats

The mouse genome sequence is derived from a female C57BL6/J animal with disease model mice (e.g. NOD, NZM) yet to be sequenced. The situation for the human genome is similar with all sequence data being derived from a very limited number of individuals. The identification of the genetic aetiology, susceptibility and protective loci or alleles of complex, human diseases will require the sequencing of many more common haplotypes. For example, a comparison of linkage hits of seven asthma studies (Altmuller et al 2001) showed 42 hits on 17 different chromosomes. When applying the significance criteria of Lander & Kruglyak (1995) six studies resulted only in suggestive linkage and one in no significant linkage.

At the transcriptional level immunomics faces the problem that inducible and cell type-specific transcripts from, for example, T cells, B cells or macrophages present at low levels in whole tissue mRNA are underrepresented in mouse cDNA (e.g. FANTOM) or human cDNA (e.g. KIAA; Ohara et al 1997) libraries. For example, cytokines are largely only induced in nucleated cells in response to danger (Schwarz et al 2001). Hence, cytokine transcripts would be anticipated to be underrepresented in libraries of unstimulated immune tissues. A framework for the construction and sharing of immunopathological cDNA library resources should be encouraged to boost the number of publicly available immune-related transcripts and transcriptional variants.

Transcriptional diversity

Antisense. Transcriptional variation will play a growing role in elucidating characteristic phenotypes in a cellular context. For example, RU2AS, the translated antisense transcript of *Ru2* (Chr 6p22.1) was found to be an HLA-B7 restricted renal carcinoma antigen (Van den Eynde et al 1999). The transcription of RU2AS is initiated by an cryptic antisense-oriented promoter in the first intron of *Ru2*. The translation of antisense is known for retroviruses (e.g. HIV) which have cytoplasmic control elements that facilitate the transport of RNA into the cytoplasm. It remains to be seen whether antisense translation in cancer cells adds to the existing range of variations shown to be caused by aberrant intron transcription (MGAT5), unspliced introns (e.g. as seen in melanoma ubiquitous mutated protein 1, Coulie et al 1995; SILV, Robbins et al 1997; and DCT, Lupetti et al 1998) or translation of alternative open reading frames (e.g. as seen in TYRP1 and CTAG1 (Wang et al 1996, 1998).

Alternative splicing. Alternative splicing, as seen in 42% of human transcripts of which 74% affect protein coding sequences (Modrek et al 2001), has relevance for non-Mendelian disease aetiology (Stamm 2002). Among immune-related transcripts, *Tcr* and *Ig* are prone to premature termination codons caused by frameshift and nonsense mutations due to rearrangement. Two groups have independently shown that aberrant *Tcrβ* transcripts with nonsense codons are down-regulated by nonsense-mediated decay factor UPF2 and internal ribosome entry sites (Wang et al 2002a, Mendell et al 2002). On the other hand, *Tcrβ* transcript mutations that generate premature termination or nonsense codons are compensated for by an increase in the number of alternatively spliced transcripts that skip the nonsense codon (Wang et al 2002b), thereby retaining some protein function. It will be interesting to see whether the interdependency of RNA scanning and alternative splicing affects TCR repertoire and is associated with disease. Nambiar and co-workers (2001) showed that the expression of alternatively spliced 3′ untranslated region of *Tcrζ* which may affect mRNA stability, was increased in patients with systemic lupus erythematosus. For bioinformatics, these findings underline the problems with trying to apply to humans, gene network modelling techniques (Friedman et al 2000) that work for bacteria and yeast, but which have less sophisticated regulatory mechanisms than mammals.

Repeats. Simple, tandem and dispersed repeats (e.g. transposable elements such as SINEs and LINEs or retrotransposed genes) can generate significant functional diversity by affecting the coding sequence. More than 15% of currently available mouse mRNA sequences with protein coding potential contain repeats that

overlap with the coding sequence (Schönbach C, Nagashima T, Silva DG, unpublished data). How much splice variation is attributable to repeats and may have disease implications is under investigation. The distribution of repeat element is not random. For example, SINEs (short interspersed nucleotide elements) occur less frequently in imprinted regions (Greally et al 2002). Depending on the location of the insertion SINE B2 may affect the transcription of neighbouring genes by acting as a pol II promoter (Ferrigno et al 2001). Endogenous retrovirus insertion (ERV) in complement *C4* affects expression of *C4A* and *C4B* by antisense inhibition (Schneider et al 2001). The TOLL-like receptor 4 cDNA (AK014533) of a C57BL6/J mouse contains a MaLR long-terminal repeat and an in-frame B2 repeat in the CDS that result in a premature termination codon and possibly a truncated (1–146 aa) form of TLR4 that lacks the cytoplasmic domain and most of the extracellular region and therefore lead to hyposensitivity to LPS (Hoshino et al 1999).

Protein motifs and domains

Motifs and domains provide a rich data source of functional clues for hypothetical proteins, signalling pathways, protein–protein interactions and regulatory or active sites. SWISS-PROT (Bairoch & Apweiler 2000), a protein knowledge base containing curated protein sequences and functional information on domains, and diseases, in 15 years grew 30-fold from 3939 entries in 1986 to 119 805 entries in release 40.36 (November 2002). Despite the human and mouse genome sequencing efforts and release of large transcriptome sets, the number of human and mouse protein sequences with curated functional information in SWISS-PROT remains low: 8855 (7.4%) for human and 5947 (5.0%) for mouse compared with an estimated proteome of about 500 000 sequences (Banks et al 2000).

The majority of sequences in the TrEMBL database of SWISS-PROT/TrEMBL or FANTOM are hypothetical proteins or otherwise uninformative sequences, with names such as 'Similar to hypothetical protein FLJ22055 . . .' or 'Hypothetical protein FLJ23636 (Similar to weakly similar to glutathione peroxidase 2′). These sequences generally have no informative homologue or common ancestor and instead match best to a non-informative homologue. Algorithms for identification of motifs are commonly used to help classify these sequences and provide functional clues on their binding and catalytic, active sites, or structure–function relations. For example 5873 of 21 050 predicted FANTOM protein sequences contain InterPro motifs/domains and for 900 (15%) sequences the InterPro name is the only functional description.

Candidate novel functions of known proteins are sometimes mediated by domains or motifs that are not found in existing databases or the literature. For

example, Kawaji et al (2002) discovered seven novel motifs using a maximum-density subgraph detection method in combination with subtraction of known motifs. Among the novel motifs was an AGPAT sub-motif containing a transmembrane domain that distinguishes mammalian 1-acyl-SN-glycerol-3-phosphate acyltransferase AGPAT3 and AGPAT4 from all other acyltransferase domain containing proteins. Whether the sub-motif plays a modulating role in inflammatory responses involving AGPAT family members remains to be shown but *in vitro* over-expression of AGPAT1 and AGPAT2 enhanced IL6 and TNFα transcription and synthesis after IL1β stimulation (West et al 1997).

Many protein functions are the consequence of cellular context plus protein structure. In turn this is dependent on the protein sequence, transcription, translation and post-translational modifications, and subcellular localization. Motif analyses will, therefore, provide only partial answers for one layer of complexity (protein sequence). The functional interpretation of motifs, particularly new motifs, requires additional efforts such as literature searching.

Using the bibliome to explore the immunome

Technological advances and large-scale initiatives are closing data gaps, but the interpretation and functional inference process is time consuming and causes a major bottleneck in result interpretation. Traditional data mining methods that find patterns occurring with high frequency are not applicable to context-dependent and interrelated data that require more detailed analysis to understand.

Immunology is a knowledge-based subject with highly descriptive presentation of results in the literature and a large body of elaborate metadata (e.g. annotations) that describe the raw data (e.g. sequences). Sequence comparison tools provide a first-pass functional inference based on similarity. If the results indicate similarity to a known sequence whose metadata are informative, subsequent analysis steps usually involve a MEDLINE search of abstracts and/or associated full-text papers. The knowledge gained from reading the abstracts or articles related to the similar sequence of interest is transferred to the query sequence. Functional knowledge extraction from microarray data or interrelating gene expression data with pathway and molecular interaction data, faces similar problems.

Ontologies

As immunomics has to deal with both complex and incomplete data, for functional inference we need to at least be able to semi-automate knowledge inference. Ontologies, such as gene ontology (GO) terms (Ashburner et al 2000) or Medical Subject Headings (Nelson et al 2001) use a curated, controlled vocabulary that link related concepts in a hierarchical structure. Given the multiplicity of functions and

context dependency of a transcript or its product, existing gene ontologies need to be refined. For example, the action of a cytokine is dependent upon the presence of cells that express the cell surface receptor for that cytokine. TGFβ1, which is expressed during B cell development, at low concentration triggers an increase in the secretion of IgG3 and IgG2, but inhibits their secretion and cell growth at higher concentration (Bouchard et al 1994, Ollila & Vihinen 2002). Current GO terms for TGFβ1 are relatively coarse grain, for example, 'extracellular matrix', 'growth factor', 'transforming growth factor-beta receptor ligand', 'cell growth', 'cell proliferation', 'defense response', 'inflammatory response', 'lymph gland development', 'myogenesis', 'necrosis', 'negative regulation of cell proliferation', 'organogenesis', 'regulation of myogenesis', and 'skeletal development'. Fine grain terms such 'B cell development' or 'regulation of isotype production', 'concentration dependent' together with MEDLINE identifier and disease MeSH terms where applicable, would improve concept mapping and establish automatic interrelations between biological process and associated diseases. At the same time retrieval of MEDLINE abstracts with GO terms and extraction of sentences that contain GO words at a given distance from the gene or protein name of interest would gain in specificity.

Text information retrieval and natural language processing

Advances in information retrieval, classification, and natural language processing have led to improved expression data analysis by literature profiling (Chaussabel & Sher 2002) or knowledge extraction tools. For example, XplorMed (Perez-Iratxeta et al 2001), MedMiner (Tanabe et al 1999) PubGene (Jennsen et al 2001) facilitate the exploration of keyword-retrieved abstracts by quantitative word dependencies and identification of co-occurrences of gene names in MEDLINE abstracts. Data mining and information extraction systems for protein–protein interactions are based on association rules (Oyama et al 2002) or Bayesian statistics (Marcotte et al 2001) and have been used to support annotation and expansion of the DIP Database of Interacting Proteins (Xenarios et al 2001). PIES (Wong 2001) and SUISEKI (Blaschke et al 2001) infer protein–protein interactions from sentences if the query word and predefined interaction words occur in the same sentence.

Interrelating text and biomolecular data

Some of the above tools lack biological context information integration, while others are specialized on the analysis of small data restricted to one topic (e.g. only protein–protein interactions). None of the tools has large-scale annotation capabilities. Annotation of free-text and computationally inferred functions is a necessity to prevent massive error propagation when inferred information is

incorporated into other curated databases. We have developed a semi-automated rule-based knowledge discovery support system with annotation capability (FACTS) that interrelates sequence-inferred molecular functional information with text-inferred functional information mined from sentences of MEDLINE abstracts, gene ontology, OMIM, BIND, DIP, motif databases and other biological databases.

Our system facilitates the intuitive exploration and annotation of mouse cDNA related molecular interactions and pathologies by simple and complex keyword or sequence searches (Nagashima et al 2003). We applied the system to nearly 28 900 cDNA clone annotations that were informative for text searches. Twenty-three per cent of cDNA clones were associated with molecular interaction-containing sentences, and 33% with gene ontology identifiers. Comparisons of sequence and text-inferred functional information with text-search informative queries revealed that three-quarters and one-quarter of transcripts shared GO terms and OMIM Morbidmap titles, assigned by both methods, respectively. The comparison of inferred disease associations by manual querying and information extraction with the semi-automated rule-based system showed that about one quarter were inferred by only one of the methods and half by both methods combined.

The protein–protein interaction networks of FACTS are shown in the context of tissue distribution or expression data, disease information (MeSH, OMIM), InterPro protein domain information and gene ontology terms. The non-canonical presentation of inferred functional associations can help visualization differences in transcriptional activity and tissue context and is therefore more amenable to analysing the complex relationships of immune molecular networks.

In a second system (GEpi) we demonstrated the inference of functional information for gene expression data during HIV1 infection of T cells (see Fig. 1) (Schönbach et al 2002). GEpi is a prototype for gene expression, epitope, protein interaction information extraction and integration. Context and temporal information is important when studying dynamic processes such as viral infection and regulation of adaptive and innate immune responses through cytokine and signal dependent transcription factors. However the use of abstract-derived information has its limitation. For example, epitope (word and sequence) information in PubMed abstracts is sparse and often lacks necessary context information, such as HLA-restriction.

Conclusions

Progressing from immunogenetics to immunomics necessitates large-scale sharing of resources and integration of huge amounts of complex data on multiple biological levels. This data must be acquired from multiple sources and integrated. Large-scale integration tools such as KLEISLI (Chung & Wong

FIG. 1. Screenshot of GEpi functional report for neurothrophin 3 (NTF3). The table shows basic information such as expression data source and profile, a summary of MEDLINE search and extraction results, as well as nomenclature, motif and pathway information integrated from external databases. Inferred molecular interaction and MeSH-based disease information are shown for NTF3 which can bind to TRKB.

1999) and CORBA that were predicted to provide the future of automated data examination and knowledge discovery for unknown reasons have not been widely adopted in academic biological research. It is too early to predict whether recent computer infrastructure sharing and data integration initiatives such as the BIOGRID or the Biomedical Research Network (BIRN) of the National Center for Research Resources at NIH will accelerate data integration in biology or immunomics. Part of the problem arises from the differing priorities and research interests of computer scientists, bioinformaticians and immunologists.

Integrated systems based on literature and biomolecular data queries that traverse biological hierarchies are required for immunomics to fulfil its promise. One-dimensional analyses at the sequence level such as peptide binding prediction

or motif discovery remain important to solve particular problems. Immunoinformatics tools that are created to support immunomics research need to focus on the multi-dimensional character and context-dependent view of immunological phenomena. In this way immunomics can facilitate the transfer of immunological knowledge to diagnosis and therapy of human immune diseases.

Acknowledgements

I thank Diego Silva and Nikolai Petrovsky for their comments.

References

Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M 2001 Genomewide scans of complex human diseases: true linkage is hard to find. Am J Hum Genet 69:936–950

Ashburner M, Ball CA, Blake JA et al 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

Arnone MI, Davidson EH 1997 The hardwiring of development: organization and function of genomic regulatory systems. Development 124:1851–1864

Bairoch A, Apweiler R 2000 The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 28:45–48

Banks RE, Dunn MJ, Hochstrasser DF et al 2000 Proteomics: new perspectives, new biomedical opportunities. Lancet 356:1749–1756

BIRN (Biomedical Information Research Network): *http://www.nbirn.net*

Blaschke C, Oliveros JC, Valencia A 2001 Mining functional information associated with expression arrays. Funct Integr Genomics 1:256–268

Bouchard C, Fridman WH, Sautes C 1994 Mechanism of inhibition of lipopolysaccharide-stimulated mouse B-cell responses by transforming growth factor-beta 1. Immunol Lett 40:105–110

Chaussabel D, Sher A 2002 Mining microarray expression data by literature profiling. Genome Biol 3:RESEARCH0055

Chung SY, Wong L 1999 Kleisli: a new tool for data integration in biology. Trends Biotechnol 17:351–355

CORBA: *http://www.corba.org*

Coulie PG, Lehmann F, Lethe B et al 1995 A mutated intron sequence codes for an antigenic peptide recognized by cytolytic T lymphocytes on a human melanoma. Proc Natl Acad Sci USA 92:7976–7980

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES 2001 High-resolution haplotype structure in the human genome. Nat Genet 29:229–232

FACTS: *http://facts.gsc.riken.go.jp/*

FANTOM1 and FANTOM2; functional annotation of mouse: *http://fantom.gsc.riken.go.jp/*

Ferrigno O, Virolle T, Djabari Z, Ortonne JP, White RJ, Aberdam D 2001 Transposable B2 SINE elements can provide mobile RNA polymerase II promoters. Nat Genet 28:77–81

Friedman N, Linial M, Nachman I, Pe'er D 2000 Using Bayesian networks to analyze expression data. J Comput Biol 7:601–620

GEpi: *http://facts.gsc.riken.go.jp/GEPI/;* UserID: guest; Password: GEPI

Greally JM 2002 Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. Proc Natl Acad Sci USA 99:327–332

HAVANA; manual curation of the human genome: *http://www.sanger.ac.uk/HGP/havana/*
Hoshino K, Takeuchi O, Kawai T et al 1999 Cutting edge: Toll-like receptor 4 (TLR4)-deficient mice are hyporesponsive to lipopolysaccharide: evidence for TLR4 as the Lps gene product. J Immunol 162:3749–3752
Jenssen TK, Laegreid A, Komorowski J, Hovig E 2001 A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 28:21–28.
Kapranov P, Cawley SE, Drenkow J et al 2002 Large-scale transcriptional activity in chromosomes 21 and 22. Science 296:916–919
Kawai J, Shinagawa A, Shibata K et al (RIKEN Genome Exploration Research Group Phase II Team and the FANTOM Consortium) 2001 Functional annotation of a full-length mouse cDNA collection. Nature 409:685–690
Kawaji H, Schönbach C, Matsuo Y et al 2002 Exploration of novel motifs derived from mouse cDNA sequences. Genome Res 12:367–378
Lander E, Kruglyak L 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247
Lander ES, Linton LM, Birren B et al 2001 Initial sequencing and analysis of the human genome. Nature 409:860–921
Lupetti R, Pisarra P, Verrecchia A et al 1998 Translation of a retained intron in tyrosinase-related protein (TRP) 2 mRNA generates a new cytotoxic T lymphocyte (CTL)-defined and shared human melanoma antigen not expressed in normal cells of the melanocytic lineage. J Exp Med 188:1005–1016
Marcotte EM, Xenarios I, Eisenberg D 2001 Mining literature for protein–protein interactions. Bioinformatics 17:359–363
Mendell JT, Ap Rhys CM, Dietz HC 2002 Separable roles for rent1/hUpf1 in altered splicing and decay of nonsense transcripts. Science 298:419–422
Modrek B, Resch A, Grasso C, Lee C 2001 Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res 29:2850–2859
Nagashima T, Silva DG, Petrovsky N et al 2003 Inferring higher functional information for RIKEN mouse full-length cDNA clones with FACTS. Genome Res 13:1520–1533
Nambiar MP, Enyedy EJ, Warke VG et al 2001 Polymorphisms/mutations of TCR-zeta-chain promoter and 3′ untranslated region and selective expression of TCR zeta-chain with an alternatively spliced 3′ untranslated region in patients with systemic lupus erythematosus. J Autoimmun 16:133–142
Nelson SJ, Johnston D, Humphreys BL 2001 Relationships in medical subject headings. In: Bean CA, Green R (eds) Relationships in the organization of knowledge. Kluwer Academic Publishers, NY, p171–184
Ohara O, Nagase T, Ishikawa K et al 1997 Construction and characterization of human brain cDNA libraries suitable for analysis of cDNA clones encoding relatively large proteins. DNA Res 4:53–59
Okazaki Y, Furuno M, Kasukawa T et al (FANTOM Consortium and RIKEN Genome Exploration Research Group Phase I & II Team) 2002 Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. Nature 420: 563–573
Ollila J, Vihinen M 2002 Microarray analysis of B-cell stimulation. Vitam Horm 64:77–99
Oyama T, Kitano K, Satou K, Ito T 2002 Extraction of knowledge on protein–protein interaction by association rule discovery. Bioinformatics 18:705–714
Perez-Iratxeta C, Bork P, Andrade MA 2001 XplorMed: a tool for exploring MEDLINE abstracts. Trends Biochem Sci 26:573–575
Rioux JD, Daly MJ, Silverberg MS et al 2001 Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. Nat Genet 29:223–228

Robbins PF, El-Gamil M, Li YF, Fitzgerald EB, Kawakami Y, Rosenberg SA 1997 The intronic region of an incompletely spliced gp100 gene transcript encodes an epitope recognized by melanoma-reactive tumor-infiltrating lymphocytes. J Immunol 159:303–308

Schwarz M, Murphy PM 2001 Kaposi's sarcoma-associated herpesvirus G protein-coupled receptor constitutively activates NF-kappa B and induces proinflammatory cytokine and chemokine production via a C-terminal signaling determinant. J Immunol 167:505–513

Shaffer AL, Rosenwald A, Hurt EM et al 2001 Signatures of the immune response. Immunity 15:375–385

Shoemaker DD, Schadt EE, Armour CD et al 2001 Experimental annotation of the human genome using microarray technology. Nature 409:922–927

Schönbach C, Nagashima T, Konagaya A, Kurochkin I 2002 Inferring protein interactions during HIV infection in context of gene expression data. Australasian Society for Immunology (ASI) Meeting, Brisbane, December 2002 (abstr)

Schneider PM, Witzel-Schlomp K, Rittner C, Zhang L 2001 The endogenous retroviral insertion in the human complement C4 gene modulates the expression of homologous genes by antisense inhibition. Immunogenetics 53:1–9

Silva DG, Schönbach C, Brusic V, Socha LA, Nagashima T, Petrovsky N 2003 Identification of novel ''pathologs'' (human disease-related gene candidates) from the RIKEN full-length mouse cDNA dataset. Genome Res 13:1559

Stamm S 2002 Signals and their transduction pathways regulating alternative splicing: a new dimension of the human genome. Hum Mol Genet 11:2409–2016

Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN 1999 MedMiner: an internet text-mining tool for biomedical information, with application to gene expression profiling. Biotechniques 27:1210–1217

Van Den Eynde BJ, Gaugler B, Probst-Kepper M et al 1999 A new antigen recognized by cytolytic T lymphocytes on a human kidney tumor results from reverse strand transcription. J Exp Med 190:1793–1800

Venter JC, Adams MD, Myers EW et al 2001 The sequence of the human genome. Science 291:1304–1351

Wang RF, Parkhurst MR, Kawakami Y, Robbins PF, Rosenberg SA 1996 Utilization of an alternative open reading frame of a normal gene in generating a novel human cancer antigen. J Exp Med 183:1131–1140

Wang RF, Johnston SL, Zeng G, Topalian SL, Schwartzentruber DJ, Rosenberg SA 1998 A breast and melanoma-shared tumor antigen: T cell responses to antigenic peptides translated from different open reading frames. J Immunol 161:3598–3606

Wang J, Vock VM, Li S, Olivas OR, Wilkinson MF 2002a A quality control pathway that down-regulates aberrant T-cell receptor (TCR) transcripts by a mechanism requiring UPF2 and translation. J Biol Chem 277:18489–18493

Wang J, Hamilton JI, Carter MS, Li S, Wilkinson MF 2002b Alternatively spliced TCR mRNA induced by disruption of reading frame. Science 297:108–110

Waterston RH, Lindblad-Toh K, Birney E et al 2002 Initial sequencing and comparative analysis of the mouse genome. Nature 2002 420:520–562

Watson JD, Crick FHC 1953 Molecular structure of nucleic acid A structure for deoxyribose nucleic acid. Nature 171:737–738

West J, Tompkins CK, Balantac N et al 1997 Cloning and expression of two human lysophosphatidic acid acyltransferase cDNAs that enhance cytokine–induced signaling responses in cells. DNA Cell Biol 16:691–701

Wong L 2001 PIES, a protein interaction extraction system. Pac Symp Biocomput 520–531

Xenarios I, Fernandez E, Salwinski L et al 2001 DIP: the database of interacting proteins: 2001 update. Nucleic Acids Res 29: 239–241

DISCUSSION

Rammensee: How would you envisage that normal scientists will be able to use this kind of tool?

Schönbach: FACTS (Functional Association/Annotation of cDNA Clones from Text/Sequence Sources), is currently being used by our collaborators through the website *http://facts.gsc.riken.go.jp/*. Perhaps Diego Silva who used FACTS for the identification and exploration of human-disease-related genes from mouse cDNAs can say something about the functionality as a 'normal' user.

Silva: FACTS is a very useful tool for a normal user, we have used it to identify novel human disease-related genes from a set of FANTOM mouse transcripts. Based on sequence analysis we identified mouse transcripts with 50–75% similarity to human genes, then using FACTS we queried MEDLINE to select by MESH term matching, those genes directly related to human disease. The final list of requested genes was then uploaded to the system for manual curation. Candidate genes were then analysed based on protein interactions, presence of motifs and genome mapping. FACTS is a very effective tool for data mining in genomics research.

Rammensee: Could I do this on your website?

Schönbach: By the time our paper (Nagashima et al 2003) has been accepted for publication the web site will be open to everyone.

Petrovsky: Essentially, what it is saying is that you have to do your own annotation. This will search the literature and give you the results, but you will then have to annotate it yourself.

Kellam: There are quite a few sources of noise that can come into this. If you started off with just the GO terms that are annotated by the literature, and if you used a carefully curated pathway with annotation to learn the natural processing and language rules, would you end up with a better predictive power from the literature?

Schönbach: Yes.

Kellam: If so, you would over-fit for what you were originally looking at.

Schönbach: Yes, this is a potential problem. This is why I didn't touch the natural language processing side. In the natural language processing community are very few attempts for using and integrating data from different domains (e.g. sequence and text data). Making toy models with highly specific or small data sets is in my opinion akin to a self-fulfilling hypothesis. For biological or biomedical applications we need a large and standardized test set (e.g. for protein–protein interactions) and be frank about the specificity of prediction. For example, when we applied to our predicted protein–protein interactions sequence-based criteria (e.g. complete sequence, sequence similarity) which are important for protein–protein interaction the specificity dropped to 5–6%.

Kellam: Does it matter if you over-learn a way of extracting information for a particular area of science?

Schönbach: It is useful for one particular area. For GO I think it will be difficult because of indirect associations and similar functions caused by different mechanisms. For retrieving subsets of molecular interactions over-learning appears to be useful. Currently we focus on protein–protein interactions. Therefore we improve the specificity for protein–protein interactions and accept the consequent loss of information on protein–DNA or protein–small molecule interactions.

Kellam: For your query construction it is really an expert system. Could you also learn a system using neural networks or genetic algorithms, rather than taking your expert system that is also inherently biased?

Schönbach: In theory. In the long-term perhaps we should use the growing body of annotated data to construct a system based on neural networks or genetic algorithms.

Petrovsky: Presumably a lot of the problems arise from inconsistent or overlapping terminology.

Schönbach: Yes. For some cytokines it is very messy. There is a cytokine called April, which is a common English word. Another example is that we analysed the top 10 hits in abstracts. 12 000 abstracts contained the word 'great', which is a synonym for G protein-coupled receptor.

Kellam: How much do errors in the literature cause problems?

Schönbach: I cannot give numbers because we have not yet evaluated of errors in the literature. But there are intrinsic problems with the usage of ambiguous symbols, inconsistent nomenclature and specificity of MeSH (Medical Subject Headings) in MEDLINE abstracts. For example MeSH can be applied for disease concept mapping. Terms such as 'Acute Disease' are very broad and a cause of false positives.

Littlejohn: How does your approach compare with other well established systems, such as Omniviz, SRS and related systems? Natural language processing and the integration of heterogeneous databases are very powerful for this.

Schönbach: I cannot comment on the systems that you mentioned. To my knowledge no one has done a fair comparison with standardized data, which is a common problem in this field. The reason I compared our work with results described by Blaschke et al (2001) is because his group also used term matching. I didn't use parsing or other more sophisticated natural language processing (NLP) techniques, so I cannot compare mine to the system you mentioned. Some of the NLP approaches that apply parsing use a very small data set for training and evaluation and obtain quite high performance. Parsing which is computationally intensive hasn't been applied to a million abstracts that we extracted, processed and

integrated in FACTS and GEpi with biomolecular data retrieved from other databases.

Littlejohn: With the annotation process, you have two options. One thing looks a lot like Lincoln Stein's DAS (distributed annotation system) system. Is it related to this or does it exploit it?

Schönbach: No, it is not related to DAS and does not exploit DAS technology. FACTS or GEpi annotation pages are accessible through a web-based interface.

Littlejohn: How are the confidence levels assigned?

Schönbach: Low level confidence values were computationally assigned using indicator words such as 'might', 'it implies', 'suggests', etc. During the annotation curators assigned confidence values 'low', 'medium' and 'high' according to annotation rules. The confidence values are qualitative and not meant to be for statistics.

Beck: How do you foresee the use of comparative data? These databases are aimed at protein–protein interactions of human proteins, but the biggest experimental data set on such interactions exist for the worm and yeast. Certainly for the worm, more than 40% of the proteins appear to have orthologues in other species, so one could extract a lot there. But if you extract the data from text-based sources, these genes might be known by totally different names. Even at the sequence level you could run into problems. There is a good experimentally generated data set but at neither the literature or the sequence level is it easy to extract out the information about the human genes.

Schönbach: We used sequence data as supporting evidence for the literature-derived data. For example we compared the text-based protein–protein interactions with records of BIND (Biomolecular Interaction Network Database, *http://www.bind.ca*), if available.

Beck: Did you go to DIP (Database of Interacting Proteins, *http://dip.doe-mbi.ucla.edu/*) first?

Schönbach: The problem with DIP database is that its founder requires users to get permission to redistribute contents, for example interacting proteins and their annotation. I have not yet obtained the permission. To answer your previous question, it depends on the sequence similarity threshold. Currently we use sequence comparison (e.g. BLASTP or FASTY). We want to be on the conservative side to decrease the number of false positives. The threshold we use is currently 90% identity.

Beck: That is very conservative. I understand why you use this, but it excludes a lot of orthologues.

Schönbach: Yes. However, term matching is ignorant to alternative splice forms, truncated or partial sequences. Since protein–protein interactions can be abolished by substitution of one amino acid residue, text-extracted protein interactions

should be carefully inspected by either sequence similarity search or reading the full-text article.

Beck: How can we overcome this problem?

Schönbach: I am looking into this issue. Another problem is that with yeast two-hybrid data. I wouldn't consider these to be hard data. This is why we introduced this annotation for the text-based data. If the annotator finds that there is a reference to yeast two-hybrid in the abstract the confidence value 'low' will be assigned.

Margalit: There are in the region of 50% false positives.

References

Blaschke C, Oliveros JC, Valencia A 2001 Mining functional information associated with expression arrays. Funct Integr Genomics 1:256–268

Nagashima T, Silva DG, Petrovsky N et al C 2003 Inferring higher functional information for RIKEN mouse full-length cDNA clones with FACTS. Genome Res 13:1520–1533

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

Mathematical models of HIV and the immune system

Dominik Wodarz

Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, MP-665, Seattle, WA 98109-1024, USA

Abstract. I describe how mathematical models have been used to elucidate the principles which govern HIV and immune system dynamics in relation to antiviral drug therapy. The review starts by introducing a basic model of virus infection and demonstrates how it was used to study HIV dynamics and to measure crucial parameters which lead to a new understanding of the disease process. Since this analysis indicates that eradication of the virus is not feasible during the lifetime of the patient, I continue to discuss mathematical models with the aim to explore how drug therapy can be used to induce long-term immunological control of the infection.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 193–215

The dynamics between viral infections and the immune system involve many different components and are multifactorial. Given such a scenario, the principles governing the dynamics and the outcome of infection cannot be understood by verbal or graphical reasoning. Mathematical models provide an essential tool to capture a set of assumptions and to follow them to their precise logical conclusions. They allow us to generate new hypotheses, suggest experiments, and to measure crucial parameters.

A particular example is HIV infection. The interactions between HIV and the immune system are more complex compared with most other infections. While immune responses have the potential to fight the virus, HIV infects $CD4^+$ T helper cells which are a central component orchestrating the generation of specific immune responses. Depending on co-receptor usage, HIV can infect other immune cells, such as macrophages and dendritic cells, which are also involved in the generation of antiviral immunity. Thus, suboptimal immune responses develop during the acute phase of the infection and can contribute to viral persistence and to the ability of the virus to mutate and evolve. The infection remains asymptomatic for years before viral load increases sufficiently and the population of $CD4^+$ T cells falls to low levels upon development of AIDS. Disease progression is associated with the evolution of specific viral

variants which are more virulent and pathogenic (e.g. evolution of strong T cell tropism, escape from immune responses, faster viral replication, and higher degrees of cytopathicity). Anti-retroviral drug therapy has been used successfully to significantly suppress viral replication and to delay disease progression in many patients. Currently, these drugs act by two mechanisms: reverse transcriptase inhibitors interfere with the process of reverse transcription and prevent the virus from infecting a cell; protease inhibitors prevent the assembly of new infectious viral particles by an infected cell. Because HIV integrates into the host genome, however, the infected cells remain unaffected and provide a viral reservoir. While most productively infected cells have a relatively short lifespan, many cells are latently infected and are very long-lived. Thus, virus eradication by drug therapy is not possible during the lifetime of the host. Because continued administration of drugs is associated with many problems, such as side effects and the generation of drug resistance, more recent research efforts have been directed at finding therapy regimes which boost HIV-specific immune responses.

In this review, I show how mathematical models can be used to understand the dynamics of HIV infection and therapy. The paper starts by describing a basic model of virus infection and continues to show how it was used to get some crucial insights into the dynamics during the asymptomatic phase of the disease. I discuss HIV and immune response dynamics during antiviral therapy and explore how drug therapy can be used to boost virus-specific immunity, resulting in long-term control of the infection.

A basic model of virus dynamics

Basic virus dynamics can be described by a model which consists of three variables (Fig. 1). The population sizes of uninfected cells, x; infected cells, y; and free virus particles, v. These quantities can either denote the total abundance within a host, or the abundance in a given volume of blood or tissue.

Free virus particles infect uninfected cells at a rate proportional to the product of their abundances, βxv. The rate constant, β, describes the efficacy of this process, including the rate at which virus particles find uninfected cells, the rate of virus entry and the rate and probability of successful infection. Infected cells produce free virus at a rate proportional to their abundance, ky. Infected cells die at a rate ay, and free virus particles are removed from the system at rate uv. Therefore, the average lifetime of an infected cell is $1/a$, whereas the average lifetime of a free virus particle is $1/u$. The total amount of virus particles produced from one infected cell, the 'burst size', is k/a.

Uninfected cells are produced at a constant rate, λ, and die at a rate dx. The average lifetime of an uninfected cell is $1/d$. In the absence of infection, the population dynamics of host cells is given by $\dot{x}=\lambda-dx$. This is a simple linear

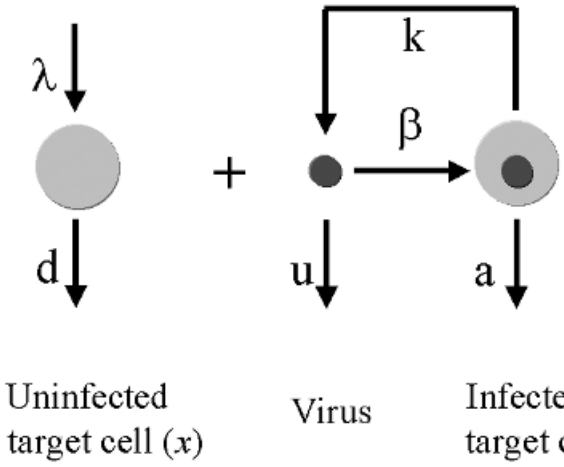


FIG. 1. Schematic illustration of the basic model of viral dynamics. Uninfected cells 'react' with free virus to give rise to infected cells; the rate constant is β. Infected cells produce free virions at a rate k. Uninfected cells, free virus and infected cells die at the rates d, u and a, respectively. Uninfected cells are replenished at a constant rate λ.

differential equation. Without virus, the abundance of uninfected cells converges to the equilibrium value λ/d.

Combining the dynamics of virus infection and host cells, we obtain a model of virus dynamics (De Boer & Perelson 1998, Nowak & Bangham 1996):

$$\begin{aligned} \dot{x} &= \lambda - dx - \beta xv \\ \dot{y} &= \beta xv - ay \\ \dot{v} &= ky - uv \end{aligned} \tag{1}$$

This is a system of non-linear differential equations. An analytic solution of the time development of the variables is not possible, but we can derive various approximations and thereby obtain a complete understanding of the system. Before infection, we have $y=0$, $v=0$, and uninfected cells are at equilibrium $x=\lambda/d$. Denote by $t=0$ the time when infection occurs. Suppose infection occurs with a certain amount of viral particles, v_0. Thus the initial conditions are $x_0=\lambda/d$, $y_0=0$, and v_0. Whether or not the virus can grow and establish an infection depends on a condition very similar to the spread of an infectious disease in a population of host individuals. The crucial quantity is the basic reproductive ratio, R_0, which is defined as the number of newly infected cells that arise from any one infected cell when almost all cells are uninfected. The rate at which one infected cell gives rise to new infected cells is given by $\beta kx/u$. If all

cells are uninfected then $x=\lambda/d$. Since the lifetime of an uninfected cell is $1/a$, we obtain $R_0=\beta\lambda k/(adu)$.

If $R_0<1$ then the virus will not spread, since every infected cell will on average produce less than one other infected cell. The chain reaction is subcritical. On average we expect $1/(1-R_0)$ rounds of infection before the virus population dies out. If on the other hand $R_0>1$, then every infected cell will on average produce more than one newly infected cell. The chain reaction will generate an explosive multiplication of virus. Virus growth will not continue indefinitely, because the supply of uninfected cells is limited. There will be a peak in viral load and subsequently damped oscillations until an equilibrium is reached. The equilibrium abundance of uninfected cells, infected cells and free virus is given by $x^*=x_0/R_0$, $y^*=(R_0-1)du/(\beta k)$, $v^*=(R_0-1)d/\beta$.

At equilibrium, any one infected cell will on average give rise to one newly infected cell. The fraction of free virus particles that manage to infect new cells is thus given by the reciprocal of the burst size, a/k. The probability that a cell (born uninfected) remains uninfected during its lifetime is $1/R_0$. The equilibrium ratio of uninfected cells before and after infection is $x_0/x^*=R_0$.

If the virus has a basic reproductive ratio much larger than one, then x^* will be greatly reduced compared to x_0, which means that during infection the equilibrium abundance of uninfected cells is much smaller than before infection. In other words, the above simple model cannot explain a situation where during a persistent viral infection almost all 'infectable' cells remain uninfected ($x^*\approx x_0$), except in the case when R_0 is only slightly bigger than unity (which is a priori unlikely in general).

Furthermore, if $R_0\gg 1$, then the equilibrium abundance of infected cells and free virus is approximately given by $y^*\approx\lambda/a$ and $v^*\approx(\lambda k)/(au)$. Interestingly, both quantities do not depend on the infection parameter β (Bonhoeffer et al 1997). The reason is that a highly infectious virus (large β) will rapidly infect uninfected cells, but at equilibrium there will only be few uninfected cells in the system. A less infectious virus (smaller β) will take longer to infect uninfected cells, but the equilibrium abundance of uninfected cells is higher. For both viruses the product βx will be the same at equilibrium, resulting in a constant rate of production of new infected cells, and therefore in similar equilibrium abundances of infected cells and free virus.

For a highly cytopathic virus (a much larger than d), the equilibrium abundance of infected cells will be small compared to the abundance of cells prior to infection. In fact, the larger a, the smaller the abundance both of infected cells and of free virus.

For a non-cytopathic virus ($a\approx d$), the equilibrium abundance of infected cells will be roughly equivalent to the total abundance of susceptible cells prior to infection.

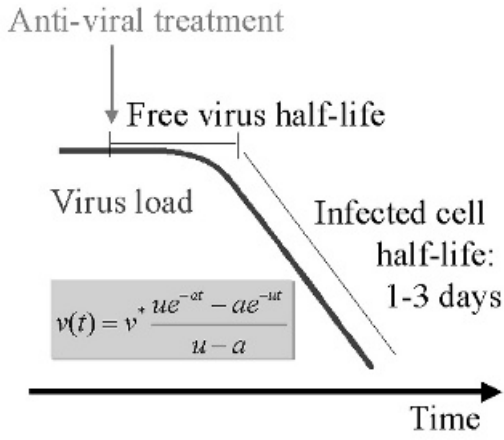


FIG. 2. Initial dynamics of HIV decay following the onset of antiviral therapy. Infected cells fall purely as an exponential function of time, whereas free virus falls exponentially after an initial shoulder phase.

Virus dynamics and antiviral therapy

During HIV infection, reverse transcriptase inhibitors prevent infection of new cells. Suppose first, for simplicity, that the drug is 100% effective and that the system is in equilibrium before the onset of treatment. Then we put $\beta=0$ in eq (1), and the subsequent dynamics of infected cells and free virus are given by $\dot{y}=-ay$ and $\dot{v}=ky-uv$. This leads to $y(t)=y^*e^{-at}$ and $v(t)=v^*(ue^{-at}-ae^{-ut})/(u-a)$ assuming $u\neq a$. Infected cells fall purely as an exponential function of time, whereas free virus falls exponentially after an initial 'shoulder phase' (Fig. 2). Since the half-life of free virus particles is significantly shorter than the half-life of virus producing cells, $u \gg a$, plasma virus abundance does not begin to fall noticeably until the end of a shoulder phase of duration $\Delta t \approx 1/u$. Thereafter virus decline moves into its asymptotic phase, falling as e^{-at}. Hence, the observed exponential decay of plasma virus reflects the half-life of virus producing cells, while the half-life of free virus particles determines the length of the shoulder phase. Note that the equation for $v(t)$ is symmetrical in a and u, and therefore if $a \gg u$ the converse is true.

In the more general case when reverse transcriptase inhibition is not 100% effective, we may replace β in eq (1) with $\bar{\beta}=s\beta$, with $s<1$ (100% inhibition corresponds to $s=0$). If the time-scale for changes in the uninfected cell abundance, $1/d$, is longer than other time-scales ($d \ll a,u$), then we may approximate $x(t)$ by x^*. It follows that the asymptotic rate of decay is

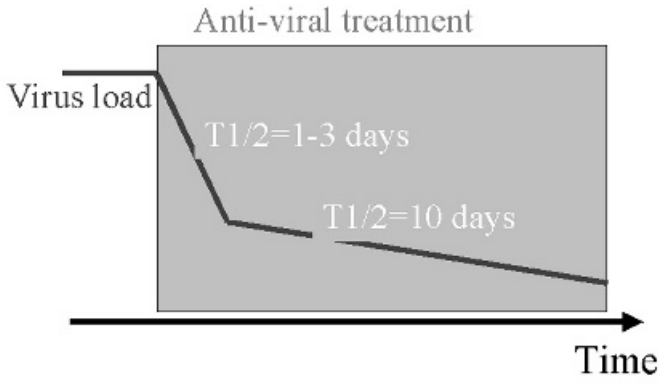


FIG. 3. Long-term dynamics of viral decay following onset of antiviral therapy. The first and rapid decay of virus (infected cell half-life of 1–3 days) is followed by a second decay phase which is significantly slower (infected cell half-life of 10 days and longer).

$\exp[-at(1-s)]$ for $u \gg a$ while the duration of the shoulder phase remains $\Delta t \approx 1/u$. Thus the observed half-life of virus producing cells, $T_{1/2}=(\ln 2)/[a(1-s)]$, depends on the efficacy of the drug.

Protease inhibitors prevent infected cells from producing infectious viral particles. Free virus particles, which have been produced before therapy starts, will for a short while continue to infect new cells, but infected cells will produce non-infectious viral particles, w. The equations become $\dot{y}=\beta xv-ay$, $\dot{v}=-uv$, $\dot{w}=ky-uw$. The situation is more complex, because the dynamics of infected cells and free virus are not decoupled from the uninfected cell population. However, we can obtain analytic insights if we again assume that the uninfected cell population remains roughly constant for the time-scale under consideration. This gives the total viral abundance as $v(t)+w(t)=v^*\ [e^{-ut}+\{(e^{-at}-e^{-ut})u/(a-u)+ate^{-ut}\}u/(a-u)]$. For $u \gg a$ this function describes a decay curve of plasma virus with an initial shoulder (of duration $\Delta t=-(2/a)\ln(1-a/u)\approx 2/u$) followed by an exponential decay of e^{-at}. The situation is very similar to reverse transcriptase inhibitor treatment. The main difference is that the viral decay function is no longer symmetrical in u and a, and therefore a formal distinction between these two rate constants is possible.

Sequential measurements of viral load in HIV1-infected patients treated with reverse transcriptase or protease inhibitors usually permit a good assessment of the slope of the exponential decline, which reflects the half-life of infected cells, $(\ln 2)/a$ (Fig. 3). This half-life is usually found to be between 1 and 3 days (Coffin 1995, Ho et al 1995, Perelson et al 1997, 1996, Wei et al 1995). The half-life of free virus particles is of the order of a few hours, possibly even less. The process that

leads to the clearance of viral particles from the peripheral blood is not understood. The half-life of virus producing cells is determined by a combination of antiviral CTL responses and viral cytopathicity (Klenerman et al 1996).

Only a small fraction of HIV1-infected cells, however, have a half-life of 2 days. These short-lived cells are thought to be productively infected $CD4^+$ T cells. They account for the production of about 99% of the plasma virus present in a patient. But most infected peripheral blood mononuclear cells (PBMCs) live much longer. During highly active anti-retroviral therapy (HAART), the relatively fast decline of plasma virus load only lasts for about one or two weeks. Subsequently the decline in viral load enters a second and slower phase (Perelson et al 1997). This second phase has a half-life of the order of 10 days (Fig. 3). The rate of decline is thought to slow down even further with time, characterized by a half-life of up to 100 days (Chun et al 1997). The population of long-lived infected cells is heterogeneous. It may comprise productively infected antigen presenting cells, such as macrophages. But more importantly, cells can become latently infected with HIV, and this population of infected cells is characterized by the longest lifespan (Chun et al 1997).

These observations have two important implications for understanding HIV infection and therapy. (i) The high turnover rate of most productively infected cells allows the virus to mutate and evolve rapidly. This could contribute to progression of the disease. (ii) While successful therapy can suppress viral load below detection limit, complete viral eradication from the patient is not possible under normal circumstances because of long-lived latently infected cells. Since lifelong therapy is not feasible (problems with compliance, resistance and side-effects), it is important to seek therapeutic strategies which result in a boost of immunity and long-term virus control in the absence of continuous therapy. This will be explored in the following section.

Using drug therapy to induce long-term immunological control

As described above, the antiviral therapy currently available cannot eradicate HIV from the host during the lifetime of the patient. Since lifelong treatment is not feasible, research has focused on identifying therapy regimes, which could result in long-term immune-mediated control of HIV in the absence of drugs. Among immune responses, cytotoxic T lymphocyte (CTL) responses have been shown to be particularly effective at fighting HIV replication (Jin et al 1999, Schmitz et al 1999). The development of protective CTL responses depends on the presence of $CD4^+$ T cell help. HIV infects and kills $CD4^+$ T cells and this can result in significant impairment of immunity against HIV. Indeed, HIV-specific helper cell impairment has been documented even in patients during the primary phase of infection (Rosenberg et al 2000).

How does this helper cell impairment influence the dynamics between HIV and specific CTL responses? In order to understand the nature of immune impairment, we have to know which immunological factors are required for efficient control of viral replication, or virus clearance. Mathematical models have identified two parameters. First, the rate of CTL activation/proliferation in response to antigen is important for limiting viral load (Nowak & Bangham 1996), and this has been shown in persistent infections such as HIV and HTLV (Jeffery et al 1999, Saah et al 1998). However, in addition virus clearance, or efficient long-term CTL-mediated control also requires antigen independent long-term persistence of memory CTLp (Wodarz et al 2000a,b). This ensures that immune pressure is maintained on the declining viral population, and this drives the virus to extinction. If CTLp are short-lived in the absence of antigen, they will decline after viral load has been reduced to low levels following CD8-mediated activity. This enables the virus to regrow, resulting in an equilibrium describing persistent viral infection in the presence of an ongoing CTL response, maintained by the persisting antigen. Hence, antigen-independent persistence of memory CTLp is required for clearance of infection. This is a new role for the antigen-independent persistence of memory CTL in viral infections.

Experiments in LCMV infected mice have shown that the development of a long-lived memory CTL response requires $CD4^+$ T cell help (Borrow et al 1996, 1998, Thomsen et al 1996, 1998). In HIV infection, the high viral load attained during the acute phase has been shown to result in the absence of significant $CD4^+$ T cell proliferative responses (Rosenberg et al 2000). This absence of $CD4^+$ T cell help could result in the failure to generate memory CTL that are long-lived in the absence of antigen. According to theory the early impairment could be the reason for persistent HIV replication and eventual loss of viral control. This hypothesis is supported by data showing that many of the CTL seen in chronic HIV infection are short-lived when viral load is reduced by drug treatment (Kalams et al 1999). This indicates that they cannot be maintained in the absence of antigenic stimulation. These CTL might be suboptimal, developing in the absence of $CD4^+$ T cell help.

These immune impairment dynamics can be captured by the following mathematical model (Wodarz & Nowak 1999).

$$\begin{aligned}
\dot{x} &= \lambda - dx - \beta xv \\
\dot{y} &= \beta xv - ay \\
\dot{v} &= ky - uv \\
\dot{w} &= cxyw - cqyw - bw \\
\dot{z} &= cqyw - bz
\end{aligned} \tag{2}$$

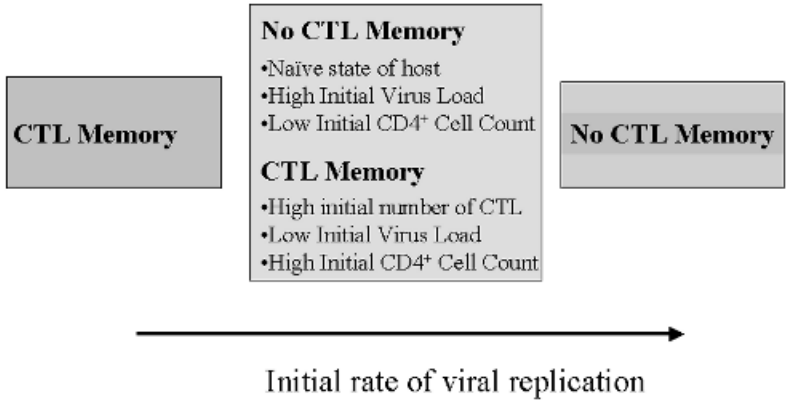


FIG. 4. Three different outcomes of the model describing CTL memory and the control of HIV. The outcome of infection depends on the replication rate of the virus. If it replicates slowly, the degree of immune impairment is low and CTL memory is generated. This corresponds to long-term non-progression. If the replication rate of the virus is high, immune impairment is strong and CTL memory can never be established, resulting in progression of disease. If the replication rate of the virus is intermediate, both outcomes are possible, depending on the initial conditions, as indicated in the diagram.

The model is based on the simple viral dynamics equations described at the beginning of this review (system 1). The target cells, x, are now assumed to be immune cells that are susceptible to HIV and that are involved in the delivery of 'help' (e.g. $CD4^+$ T cells or antigen presenting cells). In addition, we introduce a CTL response. The population of CTL is subdivided into precursors or CTLp, w, and effectors or CTLe, z. CTLp are assumed to proliferate in response to antigenic stimulation, and then to differentiate into effectors. CTLp proliferate at a rate $cxyw$ and die at a rate bw. This means, that proliferation not only requires antigen, y, but also the presence of uninfected helper cells, x. The higher the viral load, the more the uninfected helper cells become depleted, and the stronger the degree of immune impairment. Differentiation into effectors occurs at a rate $cqyw$ and is thus not assumed to require help. Finally, CTLe die at a rate bz. Thus, the mechanism of impairment underlying the model is that low levels of help result in more CTL differentiation than proliferation which eventually leads to extinction of the helper-dependent CTL response. The results do not, however, rely on this particular mechanism. The conclusions reached from this model remain qualitatively similar as long as it is assumed that high levels of viral load increase the amount of immune impairment (e.g. by alternative mechanisms such as anergy).

The behaviour of the model depends on the rate of viral replication relative to the strength of the CTL response. Three parameter regions can be distinguished (Fig. 4). (i) If the viral replication rate is slow and lies below a threshold, the degree

of immune impairment is weak and CTL memory is established. The outcome of infection is long-term control. This outcome could correspond to the long-term non-progressors. They are characterized by sustained high levels of CTL despite very low viral loads even 15–20 years after infection. (ii) If the replication rate of the virus is high and lies above a threshold, viral growth and immune impairment are overwhelming. CTL memory cannot be established and long-term virus control cannot be achieved. (iii) If the replication rate of the virus is intermediate, both outcomes of infection are possible: establishment of CTL memory leading to long-term control of HIV; and failure to establish CTL memory leading to disease progression. Which of the two outcomes is attained depends on the initial conditions, most importantly on the initial number of CTLs. If a host is naive and the initial number of specific CTLs is low, the system is likely to converge on the outcome describing failure of CTL memory and disease progression. This outcome is also promoted by high initial viral loads.

On the other hand, if the initial number of specific CTLs is high, maintenance of sustained CTL memory and long-term control is achieved. This outcome is also promoted by low initial viral loads. We assume that HIV lies in the parameter region where the outcome of infection depends on the initial conditions. In this scenario, naïve hosts fail to establish CTL memory and become progressors. However, since the CTL memory and control equilibrium is still stable, the model suggests that HAART can be used to establish CTL memory and to switch a progressor into a state of long-term non-progression. According to the model this can be done by a phase of early therapy (Fig. 5). The immune system is

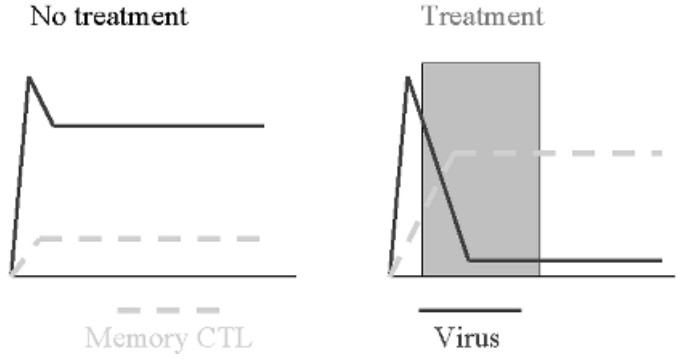


FIG. 5. Early therapy can lead to long-term control of HIV. During natural infection, HIV replicates to high levels during the asymptomatic phase of the infection. This results in high levels of immune impairment, failure to generate CTL memory, and high viral load. Early therapy prevents HIV from overwhelming the immune system while providing an antigenic stimulus. This allows the development of a CTL memory response. Once memory has been generated, it will be maintained upon cessation of treatment. Hence, early therapy can convert a progressor into a state of long-term non-progression. Experimental verification of this idea comes from SIV infection in macaques and HIV infection in humans.

provided with an antigenic stimulus, but treatment prevents the virus from reaching to high levels and significantly impairing immunity. Sufficient levels of specific $CD4^+$ T cell help are preserved, and a CTL memory response can develop. Once the memory CTL have been generated, cessation of treatment will result in maintenance of viral control. This is because the starting conditions have been altered by therapy: The initial level of memory CTL upon cessation of treatment is high.

These therapy regimes have also been studied experimentally (Lifson et al 2000, 2001). Macaques were infected with SIV, and treatment was started 24 h and 72 h post-infection. Animals that received treatment 24 h p.i. showed boosted $CD4^+$ cell proliferative responses and long-term viral control if therapy was stopped after 4 weeks. Animals that received treatment 72 h p.i. required 8 weeks of therapy to achieve improved immunological control. Animals that were characterized by undetectable viral load following cessation of treatment received a homologous re-challenge (with the same SIV isolate). Re-challenge was followed by a self-contained small blip of viraemia which was subsequently reduced below the limit of detection. Similar results were observed when the same animals were re-challenged with a more virulent SIV strain about a year after infection. When $CD8^+$ T cells were subsequently depleted with antibodies, viral load increased dramatically. These experimental results suggest that early therapy can substantially alter the dynamics between HIV and the immune system, and that sustained viral control can be achieved. They further demonstrate that protection is based on CTL responses, and that memory has been successfully generated (protection against re-challenge), as suggested by our model.

Further mathematical modelling has been used to explore the relationship between the efficacy of the drugs and the duration of therapy required for successful induction of long-term control (Komarova et al 2003). The results are summarized in Fig. 6. If therapy is strong, viral load is quickly reduced to very low levels. Upon start of therapy, the model suggests a temporary phase of CTL expansion before the response declines to insignificant levels. The reason is as follows. Upon start of treatment, sufficient antigenic stimulation is still present and immune impairment is reduced. This allows the CTL response to expand. As viral load declines to very low levels, however, the amount of antigenic stimulation is not sufficient to maintain the response during treatment, hence the decline. In order to maximize the chances of success, therapy should be stopped when the CTL response is around its highest levels. Therefore, treatment should be stopped relatively early, before the response has declined to low levels (Fig. 6a). If treatment is continued for too long, cessation of treatment will result in virus rebound to pre-treatment levels (Fig. 6a). The situation is different if antiviral therapy is less efficient (Fig. 6b). Now, viral load is reduced less efficiently during treatment. As a result, the amount of antigenic stimulation during therapy is

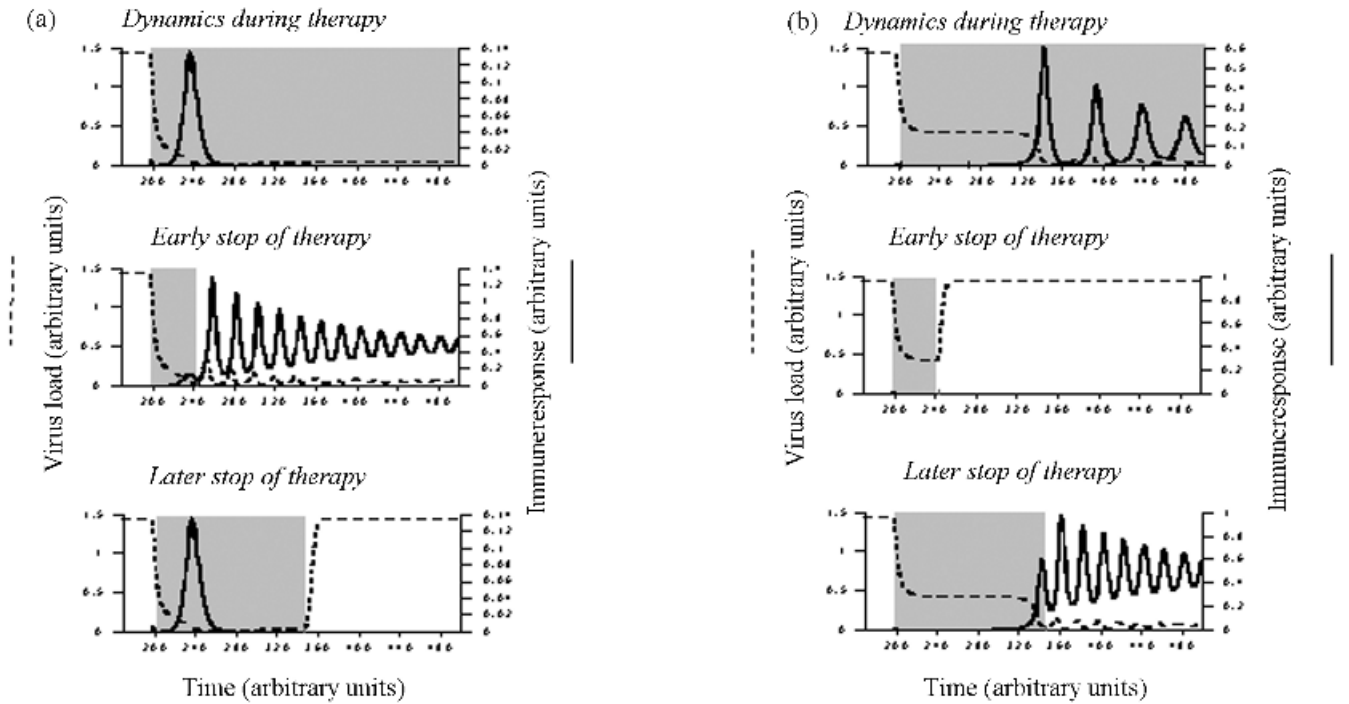


FIG. 6. Drug efficacy and timing of therapy. (a) Strong therapy reduces viral load quickly to low levels. This results in a temporary expansion of immune responses followed by a decline to insignificant levels. Immune control is achieved if therapy is stopped early before immunity has significantly declined. If therapy is continued for too long, we observe virus rebound upon cessation of treatment. (b) Weaker therapy results in expansion of immune responses after a given time delay. These responses are maintained during treatment because viral load is suppressed less efficiently and antigenic drive is still preserved. Thus, long-term control is achieved if therapy is stopped after a time threshold, once the responses have risen above a certain level. If therapy is stopped too early, we observe rebound of viral load upon cessation of treatment.

sufficient to maintain the CTL response. As shown in Fig. 6b, following a certain time delay after start of therapy, the CTL response expands and is sustained. After the response has risen above a certain level, cessation of treatment will result in long-term control. Thus, if treatment is relatively weak, therapy has to be continued beyond a time threshold before it can be stopped (Fig. 6b). If treatment is ended too early, we observe virus rebound to pre-treatment levels (Fig. 6b). The SIV experiments discussed above could correspond to a scenario where treatment is relatively weak. This is because a single drug, PMPA, has been used in these studies, and this is known to be much less efficient than the combination therapy used with HIV-infected patients.

So far, I have discussed how a single phase of drug therapy can result in the induction of long-term immunological control of HIV. Recently, so-called structured therapy interruptions (STI) have received much attention. This involves temporary interruptions of treatment and is thought to result in a boost of HIV-specific immune responses. Mathematical analysis (Komarova et al 2003), however, indicates, that in most cases interruptions are not required and that long-term control can be achieved by a single phase of therapy if the combination of drug efficacy and treatment duration is optimized. Modelling suggests that interruptions are only required if the efficacy of antiviral drugs is very strong. In this case, the temporary phase of CTL expansion upon start of treatment is not of sufficient magnitude to enable the induction of long-term control. In this case, interruptions can help to boost the response to sufficiently high levels (Komarova et al 2003).

All the treatment regimes discussed here are only likely to work if treatment is initiated relatively early in the infection process, preferably during the acute phase of the infection. As the disease progresses, HIV deletes the necessary immunological specificities. Therefore, treatment cannot be used to induce immune responses anymore. During the chronic phase of the infection, research should focus on a combination of drug therapy and therapeutic vaccination approaches to boost the level of immune responses and to convert a patient to a state of long-term non-progression. Such treatment strategies also keep the amount of viral replication at a minimum, which reduces the likelihood of generating mutations conferring immune escape or drug resistance (Bonhoeffer et al 2000).

Conclusion

This review has shown the importance of mathematical models for understanding infection dynamics, and in particular HIV dynamics. I demonstrated how a simple model of viral infection can be applied to data in order to measure crucial parameters which can lead to important new insights. I described how

mathematical models can be used to generate new hypotheses which could explain the failure of the immune system to contain HIV infection and elucidate the principles underlying immunological control. These insights were applied to guide therapy regimes aimed at long-term immune-mediated control of HIV. While some of the theoretical results have been backed up by experimental studies of SIV infected macaques, more experimental work has to be coupled with mathematical models in order to test theories in more detail and to measure more parameters.

References

Bonhoeffer S, May RM, Shaw GM, Nowak MA 1997 Virus dynamics and drug therapy. Proc Natl Acad Sci USA 94:6971–6976

Bonhoeffer S, Rembiszewski M, Ortiz GM, Nixon DF 2000 Risks and benefits of structured antiretroviral drug therapy interruptions in HIV-1 infection. Aids 14:2313–2322

Borrow P, Tishon A, Lee S et al 1996 CD40L-deficient mice show deficits in antiviral immunity and have an impaired memory CD8+ CTL response. J Exp Med 183:2129–2142

Borrow P, Tough DF, Eto D et al 1998 CD40 ligand-mediated interactions are involved in the generation of memory $CD8^+$ cytotoxic T lymphocytes (CTL) but are not required for the maintenance of CTL memory following virus infection. J Virol 72:7440–7449

Chun TW, Stuyver L, Mizell SB et al 1997 Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. Proc Natl Acad Sci USA 94:13193–13197

Coffin JM 1995 HIV population dynamics *in vivo*: implications for genetic variation, pathogenesis, and therapy. Science 267:483–489

De Boer R J, Perelson AS 1998 Target cell limited and immune control models of HIV infection: a comparison. J Theor Biol 190:201–214

Ho DD, Neumann AU, Perelson AS, Chen W, Leonard JM, Markowitz M 1995 Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature 373:123–126

Jeffery K J, Usuku K, Hall SE et al 1999 HLA alleles determine human T-lymphotropic virus-I (HTLV-I) proviral load and the risk of HTLV-I-associated myelopathy. Proc Natl Acad Sci USA 96:3848–3853

Jin X, Bauer DE, Tuttleton SE et al 1999 Dramatic rise in plasma viremia after $CD8^+$ T cell depletion in simian immunodeficiency virus-infected macaques. J Exp Med 189:991–998

Kalams SA, Goulder PJ, Shea AK et al 1999 Levels of human immunodeficiency virus type 1-specific cytotoxic T-lymphocyte effector and memory responses decline after suppression of viremia with highly active antiretroviral therapy. J Virol 73:6721–6728

Klenerman P, Phillips RE, Rinaldo CR et al 1996 Cytotoxic T lymphocytes and viral turnover in HIV type 1 infection. Proc Natl Acad Sci USA 93:15323–15328

Komarova NL, Barnes E, Klenerman P, Wodarz D 2003 Boosting immunity by antiviral drug therapy: a simple relationship among timing, efficacy, and success. Proc Natl Acad Sci USA 100:1855–1860

Lifson JD, Rossio JL, Arnaout R et al 2000 Containment of simian immunodeficiency virus infection: cellular immune responses and protection from rechallenge following transient postinoculation antiretroviral treatment. J Virol 74:2584–2593

Lifson JD, Rossio JL, Piatak M et al 2001 Role of $CD8^+$ lymphocytes in control of simian immunodeficiency virus infection and resistance to rechallenge after transient early antiretroviral treatment. J Virol 75:10187–10199

Nowak MA, Bangham CR 1996 Population dynamics of immune responses to persistent viruses. Science 272:74–79
Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD 1996 HIV-1 dynamics *in vivo*: virion clearance rate, infected cell life-span, and viral generation time. Science 271:1582–1586
Perelson AS, Essunger P, Cao Y et al 1997 Decay characteristics of HIV-1-infected compartments during combination therapy. Nature 387:188–191
Rosenberg ES, Altfeld M, Poon SH et al 2000 Immune control of HIV-1 after early treatment of acute infection. Nature 407:523–526
Saah AJ, Hoover DR, Weng S et al 1998 Association of HLA profiles with early plasma viral load, CD4+ cell count and rate of progression to AIDS following acute HIV-1 infection. Multicenter AIDS Cohort Study. AIDS 12:2107–2113
Schmitz JE, Kuroda MJ, Santra S et al 1999 Control of viremia in simian immunodeficiency virus infection by CD8+ lymphocytes. Science 283:857–860
Thomsen AR, Johansen J, Marker O, Christensen JP 1996 Exhaustion of CTL memory and recrudescence of viremia in lymphocytic choriomeningitis virus-infected MHC class II-deficient mice and B cell-deficient mice. J Immunol 157:3074–3080
Thomsen AR, Nansen A, Christensen JP, Andreasen SO, Marker O 1998 CD40 ligand is pivotal to efficient control of virus replication in mice infected with lymphocytic choriomeningitis virus. J Immunol 161:4583–4590
Wei XP, Ghosh SK, Taylor ME et al 1995 Viral dynamics in human immunodeficiency virus type 1 infection. Nature 373:117–122
Wodarz D, Nowak MA 1999 Specific therapy regimes could lead to long-term control of HIV. Proc Natl Acad Sci USA 96:14464–14469
Wodarz D, May RM, Nowak MA 2000a The role of antigen-independent persistence of memory cytotoxic lymphocytes. Int Immunol 12:467–477
Wodarz D, Page KM, Arnaout RA, Thomsen AR, Lifson JD, Nowak MA 2000b A new theory of cytotoxic T-lymphocyte memory: implications for HIV treatment. Philos Trans R Soc Lond B Biol Sci 355:329–343

DISCUSSION

Littlejohn: The shapes of the curves are pretty clear, but it is the absolute values that seem to matter. How do you know when you are at the right absolute values to intervene?

Wodarz: In practical terms I can't tell you. All I can tell you is that if you put the patient on therapy, you should monitor the immune responses. If you see that the immune response has peaked, you can try stopping therapy. Every patient will have different parameter values, and these are more qualitative results trying to understand how these dynamics work.

Littlejohn: In the clinic can you tell when someone has peaked?

Wodarz: If you monitor CTL responses you can see the fluctuations. These are not rational guidelines on how to treat patients: we are just trying to gain insights into how patients respond.

Rammensee: Where do your oscillating T cell responses come from? Is this calculated by you?

Wodarz: This is a sort of predator–prey model.

Rammensee: Doesn't experience teach us that if the T cell response goes up the virus is reduced?

Wodarz: Yes, but we are looking at persistent infection. The T cell response goes up, the virus disappears, the T cell response goes down and the virus comes up. This results in damped oscillations.

Rammensee: Are these oscillations seen in the absence of mutations?

Wodarz: Yes, they are, but in some parameter regions they aren't. I don't attribute too much significance to the presence of the oscillations themselves, more to the eventual outcome.

Silva: Have you correlated clinical data with your model?

Wodarz: The way we test the model at the moment is rather crude and qualitative. We have a SIV-infected monkey that doesn't receive the treatment and another that does, and we look at the outcomes.

De Groot: In the studies I described earlier in which we looked at the CTL epitopes in HIV patients, we noticed a certain cohort of patients that were extremely adherent to therapy with undetectable viral loads, and they had very few CTL responses. When we moved to a different patient cohort where they were less adherent and had blips of virus present in their blood, we had a lot of CTL responses to the epitopes that we were mapping. This gives clinical back-up for what you are describing. We assumed that it was this loss of immune response and it is nice to see your model predict that. It is very difficult to detect in any particular patient. You get a flat viral load. How many months out would you stop therapy? What would be your indicator for stopping treatment?

Wodarz: In an ideal scenario I would monitor as closely as possible. If there was an immune response that I thought was responsible for killing the virus, I would watch it grow up. Eventually it will go down again, and then I would try to stop just after it has peaked. This is when you have reached some maximal level of immune response.

De Groot: In strategic treatment interruption (STI) studies, in most of the patients their viral load goes right up after the STI almost as if they are naïve, and had not protective immune responses. Presumably, due to effective viral suppression by HAART you lose the immune response, and then if you take away the drugs you re-expose the 'almost-naïve' immune system to the virus.

Wodarz: That is not interruption, because you have just described one phase of treatment. What you do is take the drugs away, the virus comes up and you give the drugs again.

De Groot: Then you might have deleted the T cell clones that are specific to those epitopes.

Wodarz: If you have deleted them then you are in trouble. That is why all the monkey experiments are performed in primary infection.

De Groot: It works in primary but it doesn't work in chronic infection. At the Barcelona meeting (the World AIDS Conference, Barcelona 2002) most of the researchers agreed that we should not use STI because it doesn't work in chronic infection. In acute infection it does work, but only if you treat at a very early stage.

Wodarz: I guess you can think of therapy regimes that somehow circumvent the necessity for CD4 cells, for example by cross-ligating CD40 to stimulate dendritic cells directly without the need for CD4 cells.

De Groot: You could deliver a vaccine that enhances a broad $CD8^+$ T cell response.

Rammensee: When? During HAART therapy or afterwards?

De Groot: That's a good question. I think you would do it during HAART, while the viral load is low.

Wodarz: I would do it then because then we have no immune impairment.

Rammensee: Could you model in a situation where antigen comes in which is not infectious?

Wodarz: Yes, it would be possible to model in exactly that framework. The effect of adding non-infectious antigen would be to push the immune response above the separatrix, and when you take the drugs away, long-term immunological control could be achieved

Perelson: That protocol has been tried in therapy interruption trials where patients on HAART were vaccinated to try to boost the immune response before therapy was withdrawn (cf. Markowitz et al 2002).

Wodarz: There is a huge CTL response but for some reason they don't do anything.

De Groot: One of the problems is that you are immunizing with a clade B antigen, and you have a chronically infected person with a diverse HIV. It is probably not the same strain of HIV so you need to immunize almost with the same strain, or cross-conserved epitopes.

Bernaschi: I would like to return to the problem of mutation of different strains especially during the chronic phase. You can probably keep one or few of the strains under control. But how do you keep under control any possible mutation if there are not enough T cells free to control new strains? What we find in our simulations is that the major problem is not a reduction in number of the T cells, but a reduction in specificity of the repertoire, leading in the long term to a fatal disease.

Wodarz: The reduction in number is important as well. What we are looking at here is the primary infection. During the primary infection the virus population is relatively homogeneous. If you manage to get good control at this stage the virus is pushed to a very low level. The ability of the virus to mutate is thus greatly reduced.

De Groot: This is why STI would be very dangerous, because you are allowing the virus to mutate under immune pressure. I have always been totally opposed to this. You are forcing evolution.

Wodarz: Yes, and also under partial drug presence.

Brusic: Have you modelled the situation reflecting normal life?

Wodarz: Then you are in the chronic phase and it becomes very difficult to do that. One interesting thing we have studied is hepatitis C infection. This also seems to impair CD4 responses. We have worked with Paul Klenerman on this, and he had patients that went on one continuous drug treatment in the chronic phase of infection. He monitored their CD4 response, and saw them go up and down. By chance, some of the patients were taken off therapy when the immune response was around the peak. Once they were taken off drug treatment they had undetectable virus load. On the other hand, in patients treated for longer until the immune response was already gone, the virus came straight back when treatment was stopped. In this respect hepatitis C is a nicer scenario because it doesn't seem to kill the immune cells; it just seems to impair their responses and the specificity is still there.

Gulukota: A couple of the points that were raised earlier considering the mutation and the diversity of the virus population, could probably be treated together in your model. To do this, instead of a single virus and one set of equations for, you could couple 10 different variants of viruses and have equations for each.

Wodarz: This has been done. However, I was interested here in the very basic dynamics of immunosuppressive infection and the immune response against this infection.

Borras-Cuesta: Why do you think that you can't cure chronic infections?

Wodarz: I don't think that. Treatments involving therapeutic vaccination during treatment in the chronic infections could work. On the other hand, using drug therapy to boost immunity during chronic HIV infection is unlikely to work because the HIV-specific responses are already strongly impaired, and are unable to react

Borras-Cuesta: Unless there has been remission of CD4.

Wodarz: Of CD8. You have to cleverly boost the CD8s. CD8 cell responses require CD4 cell help. If CD4 helper cells have been depleted by the virus, you can try to directly stimulate CD8 cells with activated antigen presenting cells (APCs) without the need for CD4 cells. Normally, CD4 cells activate APCs, and APCs activate CD8 cells. You can circumvent CD4 cells by artificially activating the APCs by cross-ligation

Borras-Cuesta: You could do that. You could also use an exogenous helper peptide. So you think you can cure chronic HIV infections?

Wodarz: You can boost the immune responses and see what happens.

De Groot: You can convert the patients to long-term non-progressives. That is the goal. There is now a completely different attitude about HIV vaccines. The aim is not prophylaxis, but conversion to long-term non-progression — containment of infection.

Kellam: What do you think about comments that people make that your modelling system is using linear dynamics, and the system might have non-linearity?

Wodarz: It is not linear; it is strongly non-linear. All the models are non-linear except the very basic treatment model.

Flower: How do you code in the strength of the drug treatment?

Wodarz: By the amount by which the infection parameter β is reduced.

Flower: So is there any relationship between those changes and some property of the actual drugs you might use in treatment?

Wodarz: I guess if you just used a single drug that is not very effective, that the infection rate would be reduced less than if triple drug therapy is used.

De Groot: People look at the slope of viral load decline to evaluate the strength of the drug therapy.

Perelson: If you analyse the slope of plasma virus decline induced by drug therapy, you can show that the rate of decline depends both on the efficacy of therapy and the death rate of productively infected cells. In fact, if you look at phase I and II clinical trials where the dosage of drugs is changed, you can correlate the slope of plasma virus decline with the drug dosage (Mittler et al 2001). It is also possible to compare different dosing regimes and early slope estimates of efficacy with longer-term correlates of outcome.

Flower: So it is not just a case of saying we will modify the infection rate, or other modelling parameter, by some arbitrary amount, say by 0.5 or 0.1.

Perelson: However, we don't have an absolute measure of drug efficacies yet for HIV therapies, although we can get measures of relative efficacy on the basis of these criteria (cf. Louie et al 2003).

De Groot: What do you think about the effect of delaying therapy, which has been the new recommendation this year? We used to start treatment at a T cell count of 500, then it was 350 and now it is 250. Have you looked at how the model handles delays in therapy?

Wodarz: It depends what you want to do with treatment. If you want to boost the immune response this doesn't make any sense. If you want to avoid resistance mutation this also might be a bad thing, because the virus will have been replicating over a longer period, and the more replication cycles the higher the chance that all sorts of mutations will have been generated.

De Groot: I agree, I worry that by delaying therapy, you are increasing the number of variant viruses you'll need to immunize against or treat with

medications. There are archived T cell epitope mutations and archived drug mutations in the pool of the virus, and you are initiating treatment later.

Wodarz: Yes. On the other hand, if the patient really suffers by having treatment earlier because of side effects, it might be impossible to initiate treatment early.

De Groot: There is a clinician versus immunologist non-dialogue that is taking place here, and the clinicians have won for the time being. They say that the drugs are toxic, and people are recognizing lipidystrophies, cardiac complications and neuropathies related to the drugs. They also talk about pill burden, although the newer treatments only require one pill in the morning and one in the evening. There are certainly lipidystrophies and one would argue that they can be managed. There is an argument that the patient is being exposed to toxicities, but people don't seem to understand what is being discussed here, which is the impact on the immune system, the archiving of the mutations and the expanding diversity of the virus. Although there is no drug pressure, the immune pressure will cause the diversity to explode.

Kesmir: I don't agree fully with that. If your CD4 count is running low, then there won't be so much selection pressure on the virus. It might be that the wild-type virus is coming back again because the resistant mutants only have a selective advantage if there is a good immune response.

Wodarz: They will probably still be maintained in the population. The longer the infection is allowed to continue, the more of them will be around.

Littlejohn: Do we know the comparative fitness of the mutants versus the wild-type?

Wodarz: The mutants are less fit than the wild-type, but they archive.

Bernaschi: There is some agreement that in the case of HIV there is a subtle impact on the general homeostasis, particularly of the CD4 cells. Do you plan to include this element in your model?

Wodarz: The model looks at responses specific to the virus. I don't really look at homeostasis. It is not just the CD4 cells that I am considering; I am looking at the compartment of immune cells susceptible to HIV, which includes macrophages, CD4 cells and dendritic cells. In more elaborate models we would want to consider these populations separately and look at how they are regulated.

Littlejohn: I have a query about the stability of the steady state. Presumably it is easy to perturb. If the immune system is dampened in its steady state it could crash very quickly.

De Groot: You could have tuberculosis, which would cause explosion of the virus. This is exactly what happens.

Petrovsky: Can you extend this model to overall T cell numbers, moving away from HIV and going into a broader model? How good is this modelling if you wanted to look at the immune response to viral infections or what happens in autoimmunity with autoreactive T cells?

Wodarz: This model is describing an immune response against an infection that damages the immune response. By definition it considers a subset of viruses that might be doing something like HIV or hepatitis. Other viral infections such as influenza do not inhibit the immune response at all. You would need a different model with different assumptions for these other viruses.

Petrovsky: They should be simpler to model, then. They don't have a feedback loop.

Wodarz: Yes.

Perelson: We are modelling the acute phase of influenza infection using simple target cell limited models.

Petrovsky: What about autoimmunity?

Perelson: That's more complicated. It is not a five day infection like influenza.

Wodarz: We don't even know the reason for autoimmunity. There is great uncertainty.

Petrovsky: The issue is really one of whether we can generate models and create an artificial autoimmunity. This might help people to go back and find out whether the reality fits the model.

Perelson: Currently, there is a lot effort in looking at T cell dynamics in the context of HIV infection. If one looks at the data on the kinetics of CD4 and CD8 cells in infected and healthy individuals, one can potentially develop models of the normal regulation of the T cell compartment and also examine the effects of infection. This has been difficult because the population of T cells is heterogeneous, and involves memory and naïve cells, antigen-specific and non-specific cells, etc. There have been a number of experiments involving labelling of T cells. For example, Cliff Lane's group at the NIH has conducted pulse labelling studies with BrdU in humans even though it is somewhat carcinogenic. A safer technique is to label with a stable isotype, such as a deuterated form of glucose, where some hydrogen atoms are replaced by deuterium. The deuterated-glucose is metabolized into deoxynucleotides so every time a cell synthesizes DNA in the presence of this label it picks up deuterium. You can then isolate any cell population in the body, e.g. $CD4^+$ T cells, do mass spectrometry on its DNA and by measuring the fraction of 'heavy' DNA it has, you can determine what fraction of the cells have divided. You can also do washout experiments where you stop labelling and then monitor the kinetics of the decay of these labels, which gives an estimate of the lifetime of any labelled cell population. In collaboration with David Ho's group at the Rockefeller University we have used deuterated glucose labelling to measure the kinetics of $CD4^+$ and $CD8^+$ T cells in humans (Mohri et al 2001, Ribeiro et al 2002).

Petrovsky: Presumably you could then combine this work with tetramers. You would then have a very powerful technology.

Perelson: Yes. In mice we can combine these techniques with CSFE labelling, which gives us the whole division history of a cell. There is a lot of powerful technology available to help us answer kinetic questions in a profound way.

Rammensee: Do you see bystander T cell activation with this kind of analysis?

Perelson: We can't tell. When we do the labelling analysis we don't look at antigen specific cells because we can not harvest enough cells from humans to isolate the million or so antigen specific ones needed for analysis by mass spectroscopy. We just have enough cells to measure the label in the entire $CD4^+$ or $CD8^+$ population.

Rammensee: We have a system for getting specific T cell responses in a mouse model with tetramers. We see not only that the tetramer-positive cells are activated, but also other cells are activated which do not stain with the tetramer. These cells do not recognize the original peptide. We don't know the significance of this.

Perelson: Observations like that go very far back in immunology. Early on in the study of B cell responses in mice, people such as Alistair Cunningham found many more antigen non-specific B cells were stimulated into antibody production than antigen-specific B cells. Also there was Jonathan Sprent's work on the maintenance of T cell memory. He found that type 1 interferon stimulated by the response to any infection would stimulate memory T cells non-specifically into dividing once.

Rammensee: Dominic Wodarz, you didn't include memory cells in your model.

Wodarz: I looked at the overall CTL response. I can go into modelling memory in more detail at another stage.

Rammensee: I would assume that memory cells have a certain low degree of proliferation depending on antigen presence or absence and on the presence of the restricting MHC molecules.

Wodarz: Memory cells have a higher degree of proliferation and can be maintained at a much lower level of antigen than effector cells. There is a whole degree of confusion there: what is a memory cell? Some people have labelling definitions, while others have functional definitions. I have considered memory and done models of it. I argue that in HIV, due to the lack of $CD4^+$ cell help, the CTL response that is generated is suboptimal and not a memory response; it therefore declines at low levels of antigen when we suppress the virus by drugs. On the other hand, if we have $CD4^+$ cell help, the CTL response that develops is one that can be maintained at low levels of antigen and that is why they control it in the long term. This presence or absence of help is what we want to model in therapy or vaccination.

De Groot: The longer you delay therapy the fewer HIV-specific $CD4^+$ T helper cells are present. If you come in with therapy at a T cell count of 250 rather than 500, there probably won't be any HIV-specific T help left to rebuild the immune system.

Wodarz: Yes. One of the hypotheses is that in HIV this is the basic fault of the immune response, and the aim of therapy should be to convert such a non-memory response to a memory response.

Perelson: Gorochov and colleagues (Gorochov et al 1998) have looked at the changes in T cell repertoire with the immunoscope, which measures the length of the CDR3 region of the T cell receptor β chain. Initially when patients are put on therapy the CDR3 distributions in the T cell Vβ regions are very skewed. With continued therapy they tend to normalize, which is surprising.

De Groot: So the broad TCR repertoire comes back.

Perelson: It is not clear whether this means that the thymus is still functional or that a diversity of T cells are present at very low levels and then come back with therapy. This doesn't tell us about the functional response; it just says that the receptor repertoire looks like it is being restored. I guess we also know that in patients who have received long-term therapy that their ability to respond to opportunistic infections comes back.

De Groot: It has been shown that this does happen but you don't necessarily get the HIV-specific immune responses back.

Perelson: That is probably because we have not been able to totally control the HIV infection. We are in this funny situation where HIV, even under therapy, is probably still replicating and infecting the HIV-specific cells preferentially.

References

Gorochov G, Neumann AU, Kereveur A 1998 Perturbation of $CD4^+$ and $CD8^+$ T-cell repertoires during progression to AIDS and regulation of the $CD4^+$ repertoire during antiviral therapy. Nat Med 4:215–221

Louie M, Hogan C, Di Mascio M et al 2003 Determining the relative efficacy of highly active antiretroviral therapy. J Infect Dis 187:896–900

Markowitz M, Jin X, Hurley A et al 2002 Discontinuation of antiretroviral therapy commenced early during the course of human immunodeficiency virus type 1 infection, with or without adjunctive vaccination. J Infect Dis 186:634–643

Mittler J, Essunger P, Yuen GJ, Clendeninn N, Markowitz M, Perelson AS 2001 Short-term measures of relative efficacy predict longer-term reductions in human immunodeficiency virus type 1 RNA levels following nelfinavir monotherapy. Antimicrob Agents Chemother 45:1438–1443

Mohri H, Perelson AS, Tung K et al 2001 Increased turnover of T lymphocytes in HIV-1 infection and its reduction by antiretroviral therapy. J Exp Med 194:1277–1287

Ribeiro RM, Mohri H, Ho DD, Perelson AS 2002 *In vivo* dynamics of T cell activation, proliferation, and death in HIV-1 infection: why are $CD4^+$ but not $CD8^+$ T cells depleted? Proc Natl Acad Sci USA 99:15572–15577

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

General discussion II

Rammensee: In this general discussion, perhaps it would be interesting to discuss the HLA nomenclature issues together with the HIV database issues. The complexity is related. Is anyone able to comment on LANL, the HIV database (*http://hiv-web.lanl.gov*)?

De Groot: I use it a lot. We also use Genbank for our sequences. We don't exclusively use LANL. One of the biggest problems with LANL is annotation. When we go through the sequences looking for country-specific information, we have to sort out whether, for example, the sequence is isolated from a patient in France or is it a sequence from France. It is a linguistics problem. You have to look at sentence structure to figure this out. I am interested to know whether Christian Schönbach's approach would be able to deal with this.

Schönbach: The majority of sequences in the HIV database have some accession numbers leading to a public database such as Genbank or SWISS-PROT. The feature table or comments can be used to extract the origin of the sequence.

De Groot: This isn't always the case. A lot of information comes from comments. There might be 20 lines of comments, and a lot of time this is where we get the information from. People put the information in and they don't put it in in a particularly thought-out way. It is a problem when we are trying to categorize 65 000 HIV sequences. On a separate note, I wanted to mention that I really appreciated the HLA descriptions by Steven Marsh. The nomenclature has always been a mystery to me and now I understand.

Beck: I have one question concerning the HLA descriptions. Why were the remaining 'w's not removed?

Marsh: The 'w's are the Bw4 and Bw6 in the HLA B locus antigens. This is because they are essentially epitopes and not molecules. The HLA C antigens still keep their 'w' and so do the alleles. This is so we don't talk abut C3 and C4, and risk mixing them up with complement components which are also encoded within the MHC. The HLA-D and HLA-DP specificities keep their 'w's just because they are essentially not used anymore: they were defined by MLC and PLT cellular assays. We left them to show that they were assigned using different cellular techniques.

Gulukota: In general relational database theories it is said that your names should be stupid: they should not contain information. It seems that you have an eight character/digit reference, with each having significance. Has any thought been

given to just using dumb names, with the ancillary information being stored somewhere else?

Marsh: Different models are being used. Some people like dumb names and others like names that tell us everything that can possibly be known. We are somewhere in the middle. Ultimately, though, the important thing about a name is that it is unique. It is nice to include some information in the name, and the structure we have does this. If we use dumb names the transplant community would go crazy, and they are the main user of this database.

De Groot: Something that would be useful to me, which perhaps you already have, would be a good definition of the amino acids that form the pockets. We use a MHC handbook that describes where the pockets are when they are known. However we need more information. Do you define them in your HLA database?

Marsh: They are not on the website. We did publish a book a few years ago called the 'HLA FactsBook' (Marsh et al 1999), and they are listed there.

De Groot: That is what we use. But this isn't updated as frequently as the database. This is useful information because we are now modelling on the basis of HLA structure. We know what the amino acids are that fit by various means. We believe that there is a huge diversity in HLA, but that probably the pockets are fairly limited.

Marsh: This is something that we would like to do next with the database, giving people the opportunity to look at models of the structure so that they can determine where the polymorphisms lie and how different alleles differ in structure. This then can be related to peptide binding.

De Groot: The reason I ask is that I had this idea that we could probably go across species: among different species it is likely that the pockets are fairly conserved. We have made one attempt to do this, and were able to produce a model in which we used the pocket profile method to predict BoLA epitopes, using a data set that was defined by eluting peptides off BoLA molecules. The method we describe (in a forthcoming paper) is a combination of the two methods. This approach could then be extended using human pocket profile definitions extrapolated to other animals. This would be very useful.

Marsh: We are providing our HLA database as a model for the MHC of other species. This is something we introduced this year for the first time and is something that we intend to expand. Any of the tools that we provide for the human HLA analysis will then be directly available for the other MHCs.

De Groot: Something that would advance the field enormously would be to have the pockets defined across species. Everything we have learned for human peptide binding could then be applied.

Rammensee: Are you talking about the prediction or modelling of pockets, not based on crystal structures?

De Groot: I am talking about looking at the amino acid residues that form the pocket (in the animal analogue of HLA): if the residues that form the pockets are similar to those that line known HLA pockets, then it might be possible to define the amino acids that are likely to bind in those pockets and construct new epitope prediction tools, for animals, using this information.

Rammensee: Still, there might be influences you don't expect. The modelling might not fit the actual structure correctly.

De Groot: Then you have to test it.

Brusic: I think we can get away without analysing pockets. Pockets represent an oversimplification of the actual structure. Instead, we can use the complete positional environment of a bound peptide. This includes every amino acid in the groove that is in the proximity of peptide residues.

De Groot: You would have to superimpose the polymorphisms on the crystal structure. If we could do this it would be great. The peptide binding modelling people would love this.

Brusic: It is doable.

Kesmir: Perhaps I can say a few words about the HIV database. There is also an HIV epitope database. In many ways this reminds me of your SYFPEITHI database, because it is also a very high quality database where Bette Korber and her group are very careful in going through the literature and taking only the epitopes that they are convinced are of high quality. We have been discussing where to find data to test our prediction systems: this is definitely a place to look, because there are already a few hundred ligands there that we know will generate a good immune response. And it is not biased by A2. They have quite a broad range of MHC molecules.

Flower: But they are highly biased by the small number of proteins that they are focusing on, compared with a whole tranche of eukaryotic and microbial genomes accessible to immunology.

De Groot: They also look based on motifs. A lot of bias is introduced to databases because motif searches are used to initiate the synthesis of the peptides.

Flower: HIV has been looked at in a great deal of detail and the database is of a very high quality, but it is still very focused on a small set of motifs and proteins.

Kesmir: You are right; it is a small genome that is being looked at.

Petrovsky: My concern is that there must be 10 times as many data out there than are getting published. Most of these databases seem to be extracting their information from published detail.

Rammensee: There is a reason for this: quality control.

Petrovsky: I don't think getting something published has anything to do with quality. It is more to do with positive results and who you know. There are lots of good T cell labs out there who do a lot of studies, and unless it is topical or related to HIV, and unless they get lots of positive responses and know some nice

reviewers, the data don't get published. These are still valuable data. Can we establish a website where people can submit data that they don't publish? I don't believe that what is published represents the true amount of good quality data.

Littlejohn: There are precedents for this in the human mutation community.

Rammensee: The problem will be the quality control.

Flower: It would be a start if journals which published immunological information stipulated that before publication the data must be submitted to an appropriate database. This would be a realistic start.

De Groot: That is a great idea. The problem is that people don't want to disclose data before publication.

Flower: I envisage a situation similar to the PDB protein structural databank, with a date of release when the data become public.

Littlejohn: Coming back to what Nikolai Petrovsky was saying, there are a lot of data out there that you don't want to publish in journals. The human mutation community is a good example, because there are a lot of diagnostic labs collecting data which wouldn't be suitable for publication, but they are good data for the database.

Flower: There still has to be some validation of this exercise, otherwise there is nothing to stop people putting massive amounts of fictitious data into the database.

Littlejohn: The human mutation community does this by appointing curators to authorise data entry. It is at least a two-step process.

Brusic: Binding motifs used to be published regularly for alleles. A similar sort of publication could be established for peptides, where people deposit data and the listing is published in a journal.

Rammensee: I think *Tissue Antigens* would be interested in this sort of section.

Brusic: If we can establish a basic quality control of data, high quality data can be published regularly in a special section.

Littlejohn: The plant molecular biology community have done this for a long time. You could submit very short single page sequence-based publications.

Marsh: This is essentially what we do with the HLA database. We insist that people submit the data to the database to get a name before they can publish it. Then we give nomenclature updates every month which give people credit for finding the sequence. Some of them don't go on to publish their data, but this doesn't matter because they are there in the well curated database. We accept the data before it is published, we check it and give the sequence official names. Only then will the journals publish the sequences, once that they have been assigned a WHO official name.

Littlejohn: The microarray community have adopted the MIAME standard, and there is a lot to be learned from this, particularly as it seems to relate to the HIV database which you were saying has some problems in terms of tracking the samples.

De Groot: It is a wonderful database that is extremely accessible, but they are putting GenBank data in, and the problem here is the original GenBank entry.

Littlejohn: The MIAME group has thought about this, and we could adopt a lot of their approach wholesale.

Schönbach: Another problem with the HIV database is that there are epitope sequences that are just assigned to HLA-A2 without any allele specificity. Those should be deleted because they cannot be used for predictions or allele-based analyses.

Kesmir: They were published in that way.

De Groot: And class II epitopes are a big problem, because they are 15-mers or 20-mers and somewhere in there is your DRB101 epitope, but it is hard to extract the precise DR epitope (which is only really about 9 amino acids long) from these data.

Littlejohn: When you have something like a MIAME, you then can have MAGE, which is a XML standard. Then you can have technologies that are MAGE compliant. Now there is a database called GeneX, which is a microarray database that is MAGE compliant. The MIAME description of minimum information is just a four-page document. A lot of this relates to where the sample comes from and how it was prepared.

De Groot: What are the kinds of standards that you use for people putting entries into the HLA database?

Marsh: We have a page of standards and conditions. For example, the sequences must be sequenced in both directions. If the sequence is determined from clones they must have sequenced in multiple clones. There are limits on the minimum amount of information we require: for HLA class II sequences we want the complete exon 2 sequence, whereas for HLA class I alleles it must be both exon 2 and exon 3. We encourage people to do full genomic sequences so we have both exons and introns. These data are starting to come through. They must submit their sequence to one of the generalist databanks. If they define novel polymorphisms that have not been described before they must provide information as to how they have gone back and retyped a specific sample with newly designed probes and primers to validate that they actually have obtained the correct sequence. Most of the new HLA polymorphisms that we see are just recombinants with little motif shuffles, and these are usually picked up with the existing reagents.

De Groot: Do people have to provide the source of the sample?

Marsh: We like to know the ethnic origin and full HLA profile of the sample that has been sequenced. In an ideal world we would also like some of the sample, and in a fantastically ideal world we would actually have the facilities to go back and re-do the sequence ourselves in house and confirm its validity.

De Groot: What are the criteria for epitope selection for the HIV epitope database?

Rammensee: The criteria for our SYFPEITHI database are as follows. For the ligands, they have to be sequenced from the MHC molecule, a tough criterion. For the T cell epitope section, there should be T cell data in the publication, and the T cell should not only recognize the peptide but also the natural target which is for example the virus-infected cell. Alternatively, if it has been shown that the peptide is most likely naturally processed, this is also accepted. For example, if a mouse is immunized with a transfected cell or a protein and the mouse then produces a T cell recognizing the peptide that has been predicted, this is accepted.

De Groot: This would exclude a lot of the HIV epitope data. People use flow cell cytometry to look at overlapping peptides and all they do is look for a positive T cell response.

Rammensee: We would exclude these peptides. It is a lot of work to evaluate these data. Stephen Marsh, how many HLA sequences do you get per month?

Marsh: About 30 submissions. Some are confirmations of previously submitted data. We get at least one sequence submitted each day on average.

Rammensee: I imagine that if one opened a site for submission of HLA-associated peptides one would get 20–30 per day. We couldn't handle this.

Petrovsky: You might need to have a team of curators distributed across different countries or sites. It is still achievable.

De Groot: You could also have different tiers of epitopes depending on how they were identified.

Brusic: Yes, you could attach a level of confidence to these data.

Kesmir: The user can also decide the level of accuracy they want.

Lefranc: The big problem with this kind of approach is that you could easily get bias in the representation of the database: some people will be keen to send sequence data in as soon as they have something (even if not very valuable), and others will never send their data (even if excellent). One of the failures of GDB was when they moved for a while to direct author submission, then it was open for anyone to enter what they wanted. This led to bias.

De Groot: What would you suggest?

Lefranc: We need to have a database that is curated with each entry being checked. For the generalist databases, the situation is different. If a group is not submitting its sequence data to the generalist databases, then it should be encouraged to do so. If I receive a journal or see a paper from a group which has not submitted its data to GenBank or EMBL, I send a small message, with a copy to the editor. In contrast, for a high quality database, I think that open submission is very dangerous.

Wingender: I strongly support this idea. There can be no real high quality database without expert annotation. Whether the data are extracted manually or they are submitted, there must be manual curation. I could give some interesting examples of the sort of junk that exists in databases like GenBank. Many people are

just happy to get accession numbers so they can publish their data and they never go back and edit the entries they are responsible for. Only they can make the error corrections.

Marsh: This is the problem we found when we first started looking at the HLA sequences. We had no control. The problem is that we may realize there are lots of junk sequences in EMBL and GenBank, but many people out there using the databases don't realize this.

Wingender: What does work is when databases are serving a small community in a well defined field where there is some kind of sociological control: the people know each other. We had a good experience with a small specialized database focusing on certain aspects of chromatin structure. This worked and the community paid close attention to what they submitted.

Gulukota: Another fundamental problem that EMBL and GenBank had from the beginning is that they didn't give any thought to evidence links: how the experiment was done.

Lefranc: Personally, I strongly believe in the importance of generalist databases, even though they have their problems.

Kellam: To paraphrase Sydney Brenner, 'We have to be careful to distinguish between junk and rubbish'. We all throw out rubbish, but we also keep junk for the future: we are not sure what we will use it for but we hold on to it anyway. If you apply too rigid criteria you end up with a database which doesn't have anything like expressed sequence tags (ESTs) in it, because when they were sequenced they could be thought of as incomplete rubbish or junk. Therefore we need databases that have standards but which aren't exclusive. You can then decide afterwards whether you want to parse certain aspects of it out, or treat it all as a large database. If you don't collate it all in the first place, you have nowhere to start from.

Brusic: We need to be realistic. The role of GenBank and EMBL is to disseminate as many data as fast as possible to the research community. A certain error rate will thus be contained in the data. The specialist databases clean out most of these errors and present the data to specialized users. We need to know the limitations and confidence levels of each database to use them effectively.

Rammensee: As has occurred several times at this meeting, we have ended up in a discussion of problems in engineering and organization. This gives a good picture of our field of immunoinformatics.

Petrovsky: This suggests that we need more organization within the field, perhaps in the form of a society or working group, to tackle these problems.

Reference

Marsh SGE, Parham P, Barber LD 1999 The HLA FactsBook. Academic Press, London

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

Immunogenomics: towards a digital immune system

Stephan Beck

Wellcome Trust Sanger Institute, Hinxton Genome Campus, Cambridge CB10 1SA, UK

Abstract. One of the major differences that set apart vertebrates from non-vertebrates is the presence of a complex immune system. Over the past 400–500 million years, many novel immune genes and gene families have emerged and their products form sophisticated pathways providing protection against most pathogens. The Human Genome Project has laid the foundation to study these genes and pathways in unprecedented detail. Members of the immunoglobulin (Ig) superfamily alone were found to make up over 2% of human genes possibly constituting the largest gene family in the human genome. A subgroup of these human immune genes, those (among others) involved in antigen processing and presentation, are encoded in a single region, the major histocompatibility complex (MHC) on the short arm of chromosome 6. My laboratory has a long-standing interest in understanding the molecular organization and evolution of the MHC. To this end, we have been generating a range of MHC genomic resources that we make available in the form of maps and databases. Much of the complex data of the immune system can be reduced to binary (on/off) information that can easily be made available and analysed by bioinformatics approaches, thus contributing to better understand immune function via a 'digital immune system'.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 223–233

The free availability of the human genome draft sequence has initiated many large-scale analysis efforts and generated the need for novel analysis tools, databases and browsers (International Human Genome Sequencing Consortium, Lander et al 2001). Regions of great immunological interest such as the T cell receptor complex and the major histocompatibility complex (MHC) had been sequenced ahead of the rest of the genome (Rowen et al 1996, MHC Sequencing Consortium 1999) but, until now, could not be analysed within the higher order context of entire chromosomes or the genome as a whole. Accessing and visualizing this great wealth of diverse genome data has been (and still is) a great challenge for bioinformatics and has resulted in new developments such as the ENSEMBL genome database at *http://www.ensembl.org/* (Hubbard et al 2002) and the UCSC genome browser at *http://genome.ucsc.edu/* (Kent et al 2002).

Many of the data that we have been generating on the MHC are already available through these two public resources. The MHC is the most important genetic region in relation to common human diseases such as autoimmunity and infection. It first emerged over 400 million years ago in early vertebrates and its main function is to provide protection against pathogens. It achieves this through sophisticated antigen processing and presentation pathways that are able to recognize self from non-self. Driven by pathogen variability, the MHC is under enormous pressure to evolve and adapt quickly. Over time, it has become the most polymorphic region in the human genome with some genes (such as *HLA-B*) having over 500 alleles. However, even subtle changes in the self/non-self recognition system can lead to genetic miscommunication and result in autoimmune diseases such as diabetes, multiple sclerosis and arthritis, to name just a few. This genetic balancing act also presents a major challenge to transplant medicine where the aim is to minimize the rejection of non-self transplants while not having to compromise the patient's immune system. Our research aims to generate, integrate and provide genetic and associated data to better understand MHC biology and disease.

Results and discussion

Using both computational and experimental approaches, we have been generating detailed profiles of the MHC as illustrated in Fig. 1. In human, the MHC (known in humans as the HLA complex) is located on the short arm of chromosome 6 (6p21.3).

Genes

The initial analysis of the MHC sequence revealed 224 gene loci of which 128 were predicted to be expressed (MHC Sequencing Consortium 1999). This makes the MHC one of the most gene-dense regions of the human genome. The high number of pseudogenes is thought to facilitate the creation of new alleles through mechanisms such as gene conversion. While some genes (e.g. HLA class I and class II genes) have a known function in immunity, the function and possible involvement in immunity of many genes remains still unclear. In order to estimate the contingent of MHC genes involved in immunity, the following criteria were used to define immune function: homology to immunoglobulin domain or other immune superfamilies (based on Pfam); immune-tissue specific expression; involvement in antigen processing and presentation (histocompatibility) or inflammation; implication in regulation of expression of immune loci; and inducible by immune mediators such as interferon. According to these criteria, about 40% of the expressed MHC genes can be associated with immune function.

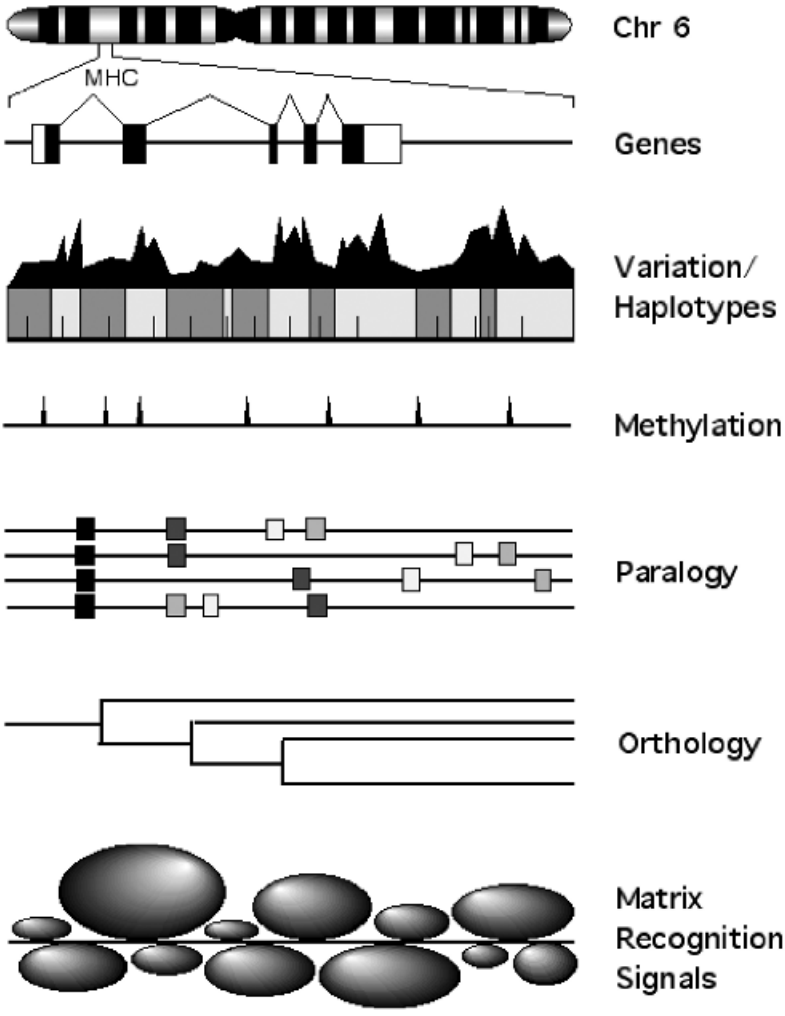


FIG. 1. A diagram illustrating the detailed profiles of the MHC at different levels that we have been generating using both computational and experimental approaches.

The clustering of so many immune-related genes in the MHC region is quite striking and unlikely to be coincidental. Another hallmark of the MHC is the extremely high linkage disequilibrium (LD) between genes suggesting that certain haplotypes are under functional selection. LD together with conserved synteny and other genomic features has contributed to re-define the classical boundaries of the MHC. In addition to the classical class I, class II and class III

subregions, extended class I and class II regions are now considered part of the extended MHC (xMHC) (Stephens et al 1999). The xMHC has recently been analysed by the Sanger Institute and the annotation is available via the 6ACE database at *http://webace.sanger.ac.uk/index.html* (Horton & Beck 2003).

Variation/haplotypes

In the quest to identify genes involved in complex disease and to better understand human evolution, over 1.4 million single nucleotide polymorphisms (SNPs) have already been mapped across the human genome (The International SNP Map Working Group, Sachidanandam et al 2001). This in turn requires an effort of millions of genotypes, a task doable in scale but not yet affordable at current cost. A possible solution to this problem is to map a representative subset of SNPs into haplotypes. Haplotypes are defined as groups of SNPs that are inherited together (e.g. are not separated by recombination). Initial studies have shown that haplotypes have an average size of around 60 kb (Reich et al 2001). In this way, one or more representative SNP(s) can be chosen from each haplotype reducing the complexity of the genotyping task accordingly. The situation in the MHC is further complicated through the finding of up to 50-fold higher variation levels than the genome average (Horton et al 1998). In order to address this problem, the MHC Haplotype Project (*http://www.sanger.ac.uk/HGP/Chr6/MHC/*) was set up to determine the complete DNA sequences of multiple common haplotypes and identify all variation (SNPs and insertions/deletions). Using this catalogue of variations, the consortium will establish the precise ancestral relationships between these haplotypes and develop a set of master SNPs suitable for the systematic identification of MHC-linked disease loci. Initially, the consortium will analyse eight haplotypes selected for their susceptibility to or their protection against type 1 diabetes and multiple sclerosis, the first two disease associations to be studied as part of the project. The resources generated as part of this project are freely available and provide a framework for association studies of all MHC-linked diseases (Allcock et al 2002).

Methylation

Disease is not only caused by genetic changes but also by epigenetic changes such as altered cytosine methylation at CpG dinucleotides (Robertson & Wolffe 2000). Epigenetics or the study of methylation patterns is one of the key areas of future research that will help to elucidate how genomes work. It combines genetics and the environment to address complex biological systems such as the plasticity of our genome. While all nucleated human cells carry the same genome, its genes are expressed differentially in time and space. Much of this is governed by epigenetic

changes resulting in differential methylation of the genome—or different epigenomes (Novik et al 2002). Individual studies over the past decades have already established the involvement of DNA methylation in imprinting, gene regulation, chromatin structure, genome stability and disease, especially cancer (Jones & Baylin 2002). With the availability of the human genome sequence, epigenetic phenomena can now be studied genome-wide giving rise to a new field, epigenomics (Novik et al 2002). The study of genome-wide methylation patterns is the aim of the Human Epigenome Consortium (Beck et al 1999). As a pilot study, the consortium is currently determining the methylation pattern of the MHC. Around 300 loci (including all expressed MHC genes) involving over 4500 CpGs will be analysed in different tissues making this pilot the largest epigenetic study to date (Novik et al 2002). Differentially methylated CpGs are identified by bisulphite sequencing and catalogued as methylation variable positions (MVPs) in the above mentioned 6ACE public database (Beck 2001). MVPs can then be epigenotyped for disease association in the same way as SNPs either by mass spectrometry (Sauer et al 2000) or microarray analysis (Adorjan et al 2002).

Paralogy

The human genome contains many (up to 10%) segmental duplications the origin of which is subject to much debate (Bailey et al 2001, 2002). As a consequence, many genes have paralogues somewhere else in the genome. Paralogues are defined as genes that have arisen by duplication from a common ancestor within the same species. While some paralogues have diverged in function over time, some have retained similar or even redundant function. The latter are of particular interest in the context of disease association studies for obvious reasons. Driven by the need to constantly evolve new alleles to recognize ever-changing pathogens, the MHC has undergone extensive duplications among other mechanisms (Beck & Trowsdale 2000). In the past, quite a few MHC paralogues have been found to cluster in three regions outside the MHC on chromosomes 1, 9 and 19 lending support to the theory of chromosome or even whole genome duplication (MHC Sequencing Consortium 1999). Again, the availability of the human genome sequence now allows us to identify and to characterize paralogous genes genome-wide with no regional bias or restrictions. Such a study is currently underway in my laboratory complemented by *in silico* and microarray-based analyses to check for possible MHC-redundant gene function outside the MHC.

Orthology

Comparative sequencing has long been recognized as a powerful approach for gene and genome analysis. It contributes to the identification of genes and other features

of interest and helps to unravel their evolutionary and phylogenetic origin. In many cases it also allows to make functional assumptions based on the common notion that conserved synteny equals conserved function. For the MHC, comparative sequencing has uncovered many interesting findings. The sequencing of the chicken MHC (also known as B-locus), for instance, has defined a minimal essential set of genes required for MHC function (Kaufman et al 1999). Compared to the human MHC, the chicken MHC was found to be about 40-fold smaller in physical size and to encode about 20-fold fewer genes. Comparative analysis of the MHC in human and mouse has further led to the formulation of the 'Framework Hypothesis' dividing MHC genes into conserved (framework) genes and non-conserved genes (Amadou 1999). By going back even further in evolution, the analysis of several fish MHCs revealed that for MHC function, MHC genes do not have to be encoded within a single complex or linkage group (Flajnik et al 1999). Today, the MHC has probably been (at least partially) sequenced in more species than any other region of the human genome, including several primates, rodents, birds, cattle, pig, cat, amphibians and fish.

Matrix recognition signals

On the DNA level, gene transcription is regulated locally by promoter and other regulatory sequences (including their epigenetic modifications) and globally by the higher-order structure of the chromatin. While many promoter sequences have been identified and catalogued in databases such as the eukaryotic promoter database (Praz et al 2002), comparatively little is known about chromatin structure. According to the inter-chromosome domain (ICD) compartment model (Cremer et al 1995), chromosomes are compartmentalized into distinct territories defined by their attachment to the nuclear matrix via short DNA sequences, called matrix attachment regions (MARs) (Boulikas 1995). MARs are thought to mediate the organization of chromatin into multiple topologically constrained loops anchored at their bases to the nuclear ribonucleoprotein matrix. In this way MARs can provide access for transcription factors and, at the same time, can provide insulation from adjacent transcription units (Volpi et al 2000). Recently, a unique bipartite sequence motif called MAR recognition signal (MRS) was identified (and experimentally verified) that can be used to predict the positions of MARs in genomic DNA sequences (van Drunen et al 1999). Such a prediction has been carried out across the MHC and, for some MARs, it could be shown that they not only bind to the nuclear matrix but also recruit the heterogenous nuclear ribonucleoprotein A1 (*hnRNP-A1*) *invivo* during transcriptional up-regulation (Donev et al 2003). *hnRNP-A1* is involved in packaging, splicing and transport of mRNA and thus confirms the proposed involvement of MARs in higher-order transcriptional control.

Conclusion

The criteria and data sets discussed above are obviously only part of the data required for the possibility to computationally simulate immune pathways. Other laboratories are already in the process of generating these missing data using chip technology, transgenetics, proteomics and, of course, bioinformatics to develop the necessary software tools towards a digital immune system.

Acknowledgements

The work described here was funded by the Wellcome Trust and the European Union.

References

Adorjan P, Distler J, Lipscher E et al 2002 Tumour class prediction and discovery by microarray-based DNA methylation analysis. Nucleic Acids Res 30:e21

Allcock RJN, Atrazhev AM, Beck S et al 2002 The MHC Haplotype project: a resource for HLA-linked association studies. Tissue Antigens 59:520–521

Amadou C 1999 Evolution of the Mhc class I region: the framework hypothesis. Immunogenetics 49:362–367

Bailey JA, Gu Z, Clark RA et al 2002 Recent segmental duplications in the human genome. Science 297:1003–1007

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE 2001 Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 11:1005–1017

Beck S 2001 Genome acrobatics: understanding complex genomes. Drug Discov Today 6:1181–1182

Beck S, Olek A, Walter J 1999 From genomics to epigenomics: a loftier view of life. Nature Biotechnol 17:1144

Beck S, Trowsdale J 2000 The human Major Histocompatibility Complex: lessons from the DNA sequence. Annual Rev Genomics Hum Genet 1:117–137

Boulikas T 1995 Chromatin domains and prediction of MAR sequences. Int Rev Cytol 162A:279–388

Cremer TS, Dietzel R, Eils P, Lichter P, Cremer C 1995 Chromosome territories, nuclear matrix filaments and inter-chromatin channels: a topological view on nuclear architecture and function. In: Brandham PE, Bennett MD (eds) Kew Chromosome Conference IV, Royal Botanic Gardens, Kew, London, UK, p 63–81

Donev R, Horton R, Beck S et al 2003 Recruitment of heterogeneous nuclear ribonucleoprotein A1 in vivo to the LMP/TAP region of the major histocompatibility complex. J Biol Chem 278:5214–5226

Flajnik MF, Ohta Y, Namikawa-Yamada C, Nonaka M 1999 Insight into the primordial MHC from studies in ectothermic vertebrates. Immunol Rev 167:59–67

Horton R, Beck S 2003 Accessing HLA sequencing data through the 6ace database. Methods Mol Biol 210:23–42

Horton R, Niblett D, Milne S et al 1998 Large-scale sequence comparisons reveal unusually high levels of variation in the HLA-DQB1 locus in the class II region of the human MHC. J Mol Biol 282:71–97

Hubbard T, Barker D, Birney E et al 2002 The Ensembl genome database project. Nucleic Acids Res 30:38–41

Jones PA, Baylin SB 2002 Fundamental role of epigenetic events in cancer. Nat Rev Genet 3:415–428

Kaufman J, Milne S, Gobel TW et al 1999 The chicken B-locus is a minimal essential major histocompatibility complex. Nature 401:923–925

Kent WJ, Sugnet CW, Furey TS et al 2002 The human genome browser at UCSC. Genome Res 12:996–1006

Lander ES, Linton LM, Birren B et al (International Human Genome Sequencing Consortium) 2001 Initial sequencing and analysis of the human genome. Nature 409:860–921

MHC Sequencing Consortium 1999 Complete sequence and gene map of a human major histocompatibility complex. Nature 401:921–923

Novik KL, Nimmrich I, Genc B et al 2002 Epigenomics: genome-wide study of methylation phenomena. Curr Issues Mol Biol 4:111–128

Praz V, Perier R, Bonnard C, Bucher P 2002 The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. Nucleic Acids Res 30:322–324

Reich DE, Cargill M, Bolk S et al 2001 Linkage disequilibrium in the human genome. Nature 411:199–204

Robertson KD, Wolffe AP 2000 DNA methylation in health and disease. Nat Rev Genet 1:11–19

Rowen L, Koop BF, Hood L 1996 The complete 685-kilobase DNA sequence of the human beta T cell receptor locus. Science 272:1755–1762

Sachidanandam R, Weissman D, Schmidt SC et al (The International SNP Map Working Group) 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409:928–933

Sauer S, Lechner D, Berlin K et al 2000 A novel procedure for efficient genotyping of single nucleotide polymorphisms. Nucleic Acids Res. 28:E13

Stephens R, Horton R, Humpray S, Rowen L, Trowsdale J, Beck S 1999 Gene organisation, sequence variation, and isochore structure at the centromeric boundary of the human MHC. J Mol Biol 291:789–799

van Drunen CM, Sewalt RG, Oosterling RW, Weisbeek PJ, Smeekens SC, van Driel R 1999 A bipartite sequence element associated with matrix/scaffold attachment regions. Nucleic Acids Res 27:2924–2930

Volpi E, Chevret E, Jones T et al 2000 Direct visualization of large chromatin loops at transcribed regions in human interphase nuclei. J Cell Sci 113:1565–1576

DISCUSSION

Rammensee: Could your modelling tell us more about the methylation in variable positions? As I understand it, methylation is highly variable. Does it make sense if you look in the different tissues? The next day it may have changed again.

Beck: It is not that variable. I gave you a simplified version of this. What we call methylation at variable positions, where there is a clear on or off, is rarely found. You have to look at a very large region, and then you get a profile. We would never just go to a single position and ask whether this is functional. In reality, we look at hundreds to thousands of CpGs to get a profile. Then we can say that within a certain cell, this is the profile. There are multiple positions that will contribute to each gene. It becomes more complicated if you look at the tissue level. Within a tissue there will be slightly different methylation profiles within cells, because there

isn't a homogeneous population of cells. It is developmentally dependent. In an ideal world, methylation data should be generated from individual, micro-dissected cells but the DNA that can be isolated from a single cell is (currently) not sufficient to analyse multiple let alone all human genes. In practice, we find that DNA isolated from macro-dissected tissue (e.g. by a pathologist) is perfectly suitable. The problem is similar to microarrays where experimental noise is filtered out by statistical analyses.

Rammensee: You said that you pool tissues of different individuals. Is this to reduce the amount of variation?

Beck: One of our biggest problems is to obtain healthy tissue. This is difficult for us to control. For the pilot project the disease we chose together with our collaborators at CNG (Evry) is psoriasis. For that we need skin biopsies from both patients and healthy individuals. Skin samples are easier to obtain than many other tissue types, which is partially why we chose it. Understandably, nobody wants to give a sample of their healthy heart, for example. Therefore, we did consider whether or not the project should be done in mouse. But, for the same reason as for the SNP project, the main and foremost interest (and therefore funding) is in human. In many ways, the mouse would be a better system to do it in because you would have all the tissues in unlimited supply. Also in respect to funding, we would like to model this project on the SNP project, and attract a combination of private and public funding. Companies are mainly interested in human methylation to develop diagnostic tools.

Silva: Is there a link between methylation patterns and disease?

Beck: The best-correlated data are those from cancer. A common scenario is that certain genes are switched off by hypermethylation, and if this includes a tumour suppressor gene a tumour can develop. Various syndromes (e.g. Beckwith–Wiedemann syndrome and Prader–Willi syndrome) have also been linked to methylation, but these are less well supported. In our approach, we like to make sure that we cover all the ground when we investigate complex diseases. Take type 1 diabetes as an example. Lots of people have been looking for variation causing diabetes in the MHC and the time has come to use a brute force approach to systematically analyse all genetic variations and consider epigenetic variations as well. It is well established, for instance, that epigenetic down-regulation can lead to an imbalance of transcripts and result in disease.

Silva: If a gene is switched off by methylation, can it be switched on again?

Beck: Theoretically yes, although little is known about this process. Methylation can be lost (and gene activity potentially be regained) if methylation is not re-established on the newly synthesized strand after replication. Usually, however, it works the other way round. Over time, the epigenome gains rather than loses methylation as part of the normal ageing process. Environmental factors such as diet can also play a role.

Littlejohn: How does the array-based detection of methylation work?

Beck: The trick is in using bisulfite treatment which converts non-methylated (but not methylated) cytosines into uracils which, in turn, become thymines after PCR amplification. Complimentary oligos for methylated and non-methylated target sequences are attached to the chip and hybridized with bisulfite treated and non-treated patient DNA (e.g. Adorjan et al 2002). Another approach to type methylation is based on mass spectrometry which is being developed by our collaborators at CNG, the French centre for genotyping in Evry.

Margalit: How does it work?

Beck: Matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS) is based on gas phase separation of biomolecules according to molecular mass. The measured difference in molecular mass is indicative whether or not the analyte was methylated.

Wingender: There is a company we are collaborating with, Epigenomics in Berlin, who are specifically working on methylation patterns.

Beck: As I mentioned before, Epigenomics and CNG are members of the consortium. Epigenomics is the leading company in the field of methylation and they are making a big contribution to this work. They came up with the chip analysis and bisulfite conversion method used in this project.

Kellam: I was interested in the looping out of the MHC and the exposure to interferon inducers. Some of these interferon-responsive genes must be constitutively expressed. In this case, before the interferon are they looped out already and are just packed tightly to the chromatin? Also, did you see looping out of the orthologous or paralogous regions that you have been mapping?

Beck: The looping out occurs very fast, about 10 mins after the addition of interferon. and smaller, not detectable sections could loop out even faster. We haven't looked at MHC orthologous or paralogous regions but did look at other regions outside the MHC as control regions.

Silva: How easy is it to scan the whole genome to look for looping out?

Beck: Now that we have a clone tile path for the entire genome, it is technically possible but it would be an enormous amount of work. It would have to be done in multiple cell lines. There are many more matrix attachment regions than are being used in a particular cell type. It is somehow similar to promoters: not every cell type produces all transcription factors. What we currently don't know is which MARs are used in which cell type. There are probably different MARs being used, for example, in B cells than in fibroblasts.

Lefranc: Can you do these kinds of things for the adaptive immune response? Would it be useful to compare the looping out of the T cell receptor and MHC?

Beck: This would be an interesting experiment to do.

De Groot: I understand that you can see the induction of gene expression by interferon γ, using the method you have developed. When you are talking about

the digital immune system what you are inferring is that it might be possible to see every step of cellular activation in some way. So you could put together a movie of what is happening in the cell at each stage.

Beck: In addition to computational simulation, this is very much my vision of a digital immune system. Another method to digitize cellular processes is by using green fluorescent protein (GFP) tagging.

Flower: Are people working towards decreasing the resolution of this process, or have we reached a fundamental limit?

Beck: I don't think the limit of resolution has been reached yet. What would be really nice to visualize in future is not only the transcriptional activation in the form of looping out but the actual RNA being made.

Kellam: There is a very similar story in B cell activation domains, where there is looping out of regions of chromatin into regions of the nucleus that are more dense with transcription factors and processing factors. When the state of the cell is changed, multiple things come off different chromosomes that all end up in a substructure of the nucleus which is more transcriptionally active.

Beck: I think you are probably referring to the inter-chromosomal domain (ICD) compartment model developed by the Cremer brothers.

Kellam: So you should be able to correlate genome-wide transcriptional patterns with the chromosome loops popping out.

Beck: If you have knowledge of which region responds to which cytokine, you could probably paint the corresponding chromosomes and stain the target regions with the same or different fluorophores to visualize multiple transcriptionally active regions.

Kellam: Does this make a prediction that complex transcriptional changes should have their genes fairly contiguous on chromosome locations?

Beck: I think so. I don't think that the chromosomal locations of genes are random. The MHC is a good example.

Littlejohn: Has anyone done something resembling an EST experiment, where S1 nuclease is used and all the fragments that are subsequently released are sequenced? This would help us to look at the regions that are looping out.

Beck: Not that I am aware of.

Reference

Adorjan P, Distler J, Lipscher E et al 2002 Tumour class prediction and discovery by microarray-based DNA methylation analysis. Nucleic Acids Res 30:e21

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

Viral bioinformatics: computational views of host and pathogen

Paul Kellam, Ria Holzerlandt, Eva Gramoustianou*[1], Richard Jenner and Antonia Kwan

*Virus Genomics and Bioinformatics Group, Department of Immunology and Molecular Pathology and *Department of Virology, University College, Windeyer Institute of Medical Sciences, 46 Cleveland Street, London W1T 4JF, UK*

Abstract. Wherever cellular life occurs, viruses are also found. As a result, complex organism and cellular antiviral responses co-evolve with virally encoded countermeasures. Since viruses co-opt or interfere with specific cellular pathways during their replication, knowledge of viral genome sequences has helped fundamental understanding of host biology. During viral infection, shifts in the balance between host and viral biological processes result in acute or chronic viral disease pathology accompanied with either active viral replication, viral containment/persistence or viral clearance. Studying host–virus interactions at the level of single gene effects, however, fails to produce a global systems-level understanding. This should now be achievable in the context of complete host and pathogen genome sequences. New experimental methods and computational approaches are rapidly developing, allowing global views of dynamic viral and cellular molecular mechanisms. Systems level virology using DNA microarrays and specific viral data resources will reveal the detailed cellular context in which viruses replicate, highlighting common and distinct antiviral mechanisms, the effect of different host cell gene expression programs, and the response of cells to similar or diverse virus types. Ultimately, microbiology and immunology will tend towards a systems-level view of how host and pathogen interact.

2003 Immunoinformatics: bioinformatic strategies for better understanding of immune function. Wiley, Chichester (Novartis Foundation Symposium 254) p 234–249

Post-genomic virology

Biological science has developed into many distinct disciplines. Now, such scientific disciplines face interpreting their knowledge base in the context of finite and defined complete genomic sequence data. This requires thinking of and using information outside the confines of previously structured research boundaries. Due to the diverse nature and large volumes of information available computational biology will play an ever-increasing role in such research.

[1]Present address: Department of Virology, Royal Free Campus, University College, London, UK.

Currently this manifests itself in gene, protein and biological function orientated databases designed to encompass a corpus of knowledge. The explosion of such data resources is charted each year in the database issue of *Nucleic Acids Research*. But, rather than these resources being dull and lifeless, they actually provide a platform for thinking genomically. For example, as viruses infect every domain of life, insights into antiviral strategies are available from bacteria to humans. This has been highlighted recently by the rise in interest in 'innate' antiviral responses such as RNA interference of plants and animals (Hannon 2002) and Toll-like receptors of insects and mammals (O'Neill 2002).

Microbes effectively co-evolve with other organisms present in their environment. As host immune complexity evolves so pathogens co-evolve, and vice versa, in a Darwinian manner. Immunoinformatics should therefore encompass pathogen bioinformatics. Post genomic research therefore allows us to take an overarching view of how a finite human genome facilitates and maintains essential biological functions and homeostasis whilst guarding against and responding appropriately to acute, persistent or commensal infections (Kellam 2001).

Viral bioinformatics

Comparative virology has documented differences and similarities in virus structures, genome types and replication strategies. Computational biology is now expanding on this theme (Kellam & Albà 2002). The starting point for functional genomics/bioinformatics is often the entire genome sequence of the organism in question. Such 'genome sequencing projects' have produced an explosion of specialist 'organism-centred' genome databases. The number and diversity of completely sequenced viral genomes in GenBank, however, far outstrips any other genome centred resource. By organizing virus genome sequences and grouping viral proteins into families of proteins that share amino acid sequence similarity a wide range of comparative virology is possible. VIDA, an animal virus database, is a prototype of such a viral centred genome data repository. VIDA organizes information based upon viral open reading frames (ORFs) from complete or partial genomic sequences derived from GenBank (Albà et al 2001a). The families within VIDA are automatically derived for all ORFs from a given virus family, for example the herpesviruses, based on conserved regions of amino acid similarity that define a viral homologous protein family (HPF). Viral ORFs can exhibit high mutation rates and can diverge quickly. Therefore, the identification of such conserved sequence regions is a valuable tool in identifying functionally important protein regions and to cross-compare different virus genomes (Albà et al 2001b, Montague & Hutchison 2000).

Perhaps more importantly, as viruses are obligate intracellular parasites and utilize many normal cellular pathways and components during their replication cycle, bioinformatics strategies can be used to identify virus proteins that interfere with the host system. A variety of methods are available to search for homologues in the host genome. Simple searches of a single protein against a database of other proteins will identify close homologues that have global or local areas of amino acid similarity. More sensitive searches can be performed against the same protein databases using conserved motifs. This relies on the fact that related proteins performing related functions have conserved regions of amino acid similarity in certain domains of the proteins (i.e. active sites). By comparing many different related protein domains, it is possible to extract a consensus motif and represent this as a Position Sequence Scoring Matrix (PSSM), which is essentially a normalized frequency of occurrence table for each amino acid position within the matrix. Searching protein databases with such PSSMs identifies more distant homologues to the group of proteins that comprise the defined motif. This method was used successfully to identify certain herpesvirus proteins as inhibitors of Fas-mediated apoptosis (Thome et al 1997) and when performed systematically identifies more host cell relatives of herpesvirus proteins in the complete human genome sequence (Holzerlandt et al 2002).

This analysis can lead to functional insights into host and pathogen processes. For example, the Kaposi's sarcoma-associated herpesvirus (KSHV; human herpesviruses 8, HHV8) encodes two proteins K3 and K5 shown to promote the down-regulation of cell surface proteins. In particular K3 promotes the endocytic down-regulation of major histocompatability complex (MHC) class I (HLA-A, -B, -C, and -E) cell surface expression by increasing the rate of endocytosis and targeting the internalized proteins for degradation. Although the precise mechanisms remain to be determined it seems clear that K3 and K5 ubiquitinate their target proteins facilitating their endocytosis and then target the internalized proteins to the lysosome in a ubiquitin-proteasome-dependent manner (Lorenzo et al 2002, Means et al 2002). The equivalent protein in murine herpesvirus 68 (MHV68), namely MK3 (ORF 12) also down-regulates murine MHC-I (H-2D) by binding to H-2D in the endoplasmic reticulum, targeting the proteins for degradation and thereby preventing cell surface expression (Boname & Stevenson 2001). The K3/K5 protein family in VIDA contains homologous viral proteins from other herpesviruses, namely, IE1 in bovine herpesvirus 4 (BHV4), MK3 (MHV68), and ORF 12 in saimirine herpesvirus 2 (HVS2), all of which contain the sequence motif known as the BKS (BHV-4, KSHV, and Swinepox) motif, a member of the PHD/LAP zinc finger class (C4HC3), but clearly differing from PHD/LAP zinc fingers due to its distinct spacing of the cysteine/histidine residues (Fig. 1). Many proteins that contain a PHD/LAP motif have been

shown to act as ubiquitin ligases but the effect of the small insertion in the BKS motif remains to be determined.

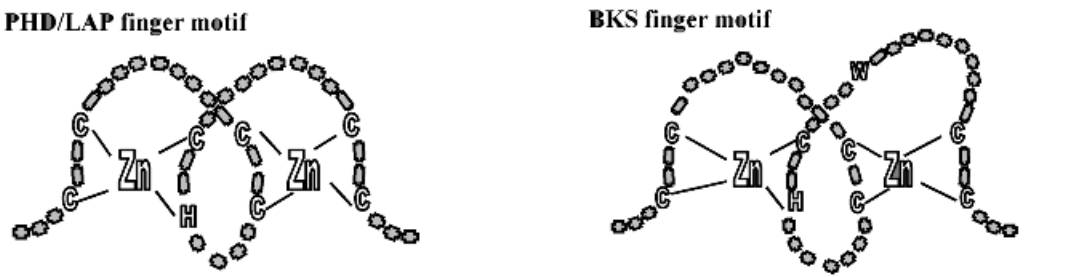
Six unannotated human proteins were identified using the BKS PSSM, all containing the highly conserved BKS finger motif. In the herpesvirus proteins the motif is always found in the N-terminus but in one human protein it appeared in the central part of the peptide whilst in another, the counterpart of murine axotrophin, at the C-terminus (Fig. 1) (Holzerlandt et al 2002, Jenner & Boshoff 2002). Interestingly, the K5 protein of KSHV, in addition to removing HLA-A and -B from the cell surface also reduces the level of cell surface, expression of intercellular adhesion molecule 1 (ICAM1) and the co-stimulatory molecule B7-2 (Coscoy & Ganem 2001, Ishido et al 2000). This suggests the possibility that the entire family of host and viral BKS motif proteins can selectively remove specific cell surface proteins.

Computational approaches alone can provide insights, as described above, but often fail to reveal related mechanisms employed by distinct viral types. Cross-comparison between any species is particularly vulnerable to misinterpretation if there is no universal understanding of the functional definitions for homologous genes. This has resulted in the Gene Ontology Consortium developing a standardized controlled vocabulary to define 'biological process', 'molecular function' and 'cellular component' for the annotation of any gene product (Ashburner et al 2000). Each of these ontologies is represented by directed acyclic graphs (DAGs); each node of the graph can have multiple 'children', in which case it is termed a parent, and each child can likewise have more than one parent (Fig. 2), thereby distinguishing DAGs from classical hierarchies. Each node of the DAG represents a defined level of function, with each child (i.e. further down the DAG) being a more specialized sub-section of the higher parent's function. Such a system is readily expanded to include viral-specific terms (R. Holzerlandt, personal communication). This should eventually lead to an integrated view of related processes encoded by different viral species and how in some case these integrate with host cell functions. For example, viruses encode extensive mechanisms for immune evasion (Alcami & Koszinowski 2000). As described, herpesviruses can interfere with MHC class I cell surface presentation, but many other diverse virus types also use this strategy. Representing this information within a DAG not only makes analysis computationally tractable but leads to a continuously updateable cross-species, information resource (Fig. 2) (Table 1).

Funtional genomics of host and pathogen

As viruses are obligate intracellular parasites, a range of regulated host-cell factors and pathways are used by viruses to enable viral gene expression and replication.

a)


PHD/LAP finger motif
BKS finger motif
Zn
Zn
Zn
Zn
H
H
W
Cx2-Cx4-11-Cx2-Cx4-Hx2-Cx9-14-Cx2-C
Cx2-Cx10-15-Cx-Cx7-Hx2-Cx3-Wx8-21-Cx2-C

b)

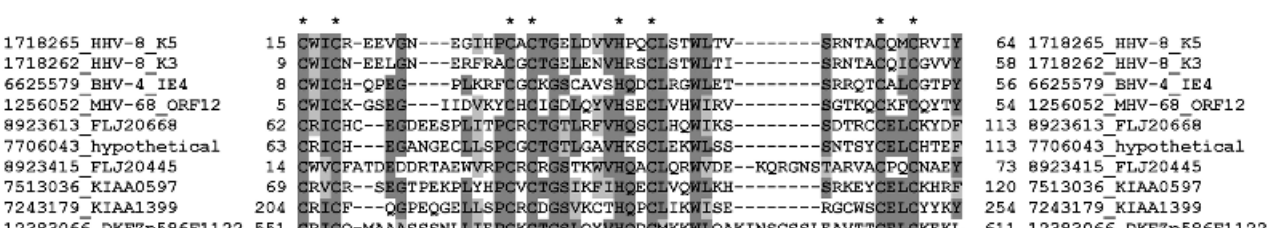

1718265_HHV-8_K5 15 CWICR-EEVGN---EGIHPCACTGELDVVHPQCLSTWLTV-------- SRNTACQMCRVIY 64 1718265_HHV-8_K5
1718262_HHV-8_K3 9 CWICN-EELGN---ERFRACGCTGELENVHRSCLSTWLTI-------- SRNTACQICGVVY 58 1718262_HHV-8_K3
6625579_BHV-4_IE4 8 CWICH-QPEG----PLKRFCGCKGSCAVSHQDCLRGWLET-------- SRRQTCALCGTPY 56 6625579_BHV-4_IE4
1256052_MHV-68_ORF12 5 CWICK-GSEG---IIDVKYCHCIGDLQYVHSECLVHWIRV-------- SGTKQCKFCQYTY 54 1256052_MHV-68_ORF12
8923613_FLJ20668 62 CRICHC--EGDEESPLITPCRCTGTLRFVHQSCLHQWIKS-------- SDTRCCELCKYDF 113 8923613_FLJ20668
7706043_hypothetical 63 CRICH---EGANGECLLSPCGCTGTLGAVHKSCLEKWLSS-------- SNTSYCELCHTEF 113 7706043_hypothetical
8923415_FLJ20445 14 CWVCFATDEDDRTAEWVRPCRCRGSTKWVHQACLQRWVDE--KQRGNSTARVACPQCNAEY 73 8923415_FLJ20445
7513036_KIAA0597 69 CRVCR--SEGTPEKPLYHPCVCTGSIKFIHQECLVQWLKH-------- SRKEYCELCKHRF 120 7513036_KIAA0597
7243179_KIAA1399 204 CRICF---QGPEQGELLSPCRCDGSVKCTHQPCLIKWISE-------- RGCWSCELCYYKY 254 7243179_KIAA1399
12383066_DKFZp586F1122 551 CRICQ-MAAASSSNLLIEPCKCTGSLQYVHQDCMKKWLQAKINSGSSLEAVTTCELCKEKL 611 12383066_DKFZp586F1122

FIG. 1. Identification of host proteins that contain the BKS finger motif found in herpesviruses and poxviruses. (a) The BKS finger motif is distinguished from the related PHD/LAP finger motif by having a conserved tryptophan residue and amino acid insertions between the fifth and six conserved cysteine residues. (b) Conservation of the BKS specific amino acids between the viral and host proteins, asterisks indicate the positions of the conserved cystiene residues found in both the PHD/LAP finger and BKS finger motifs. (c) Schematic representation of the relative positions of the BKS finger motifs in the viral and host proteins. GenBank accession numbers are provided for each protein and total protein amino acid length is indicated in brackets

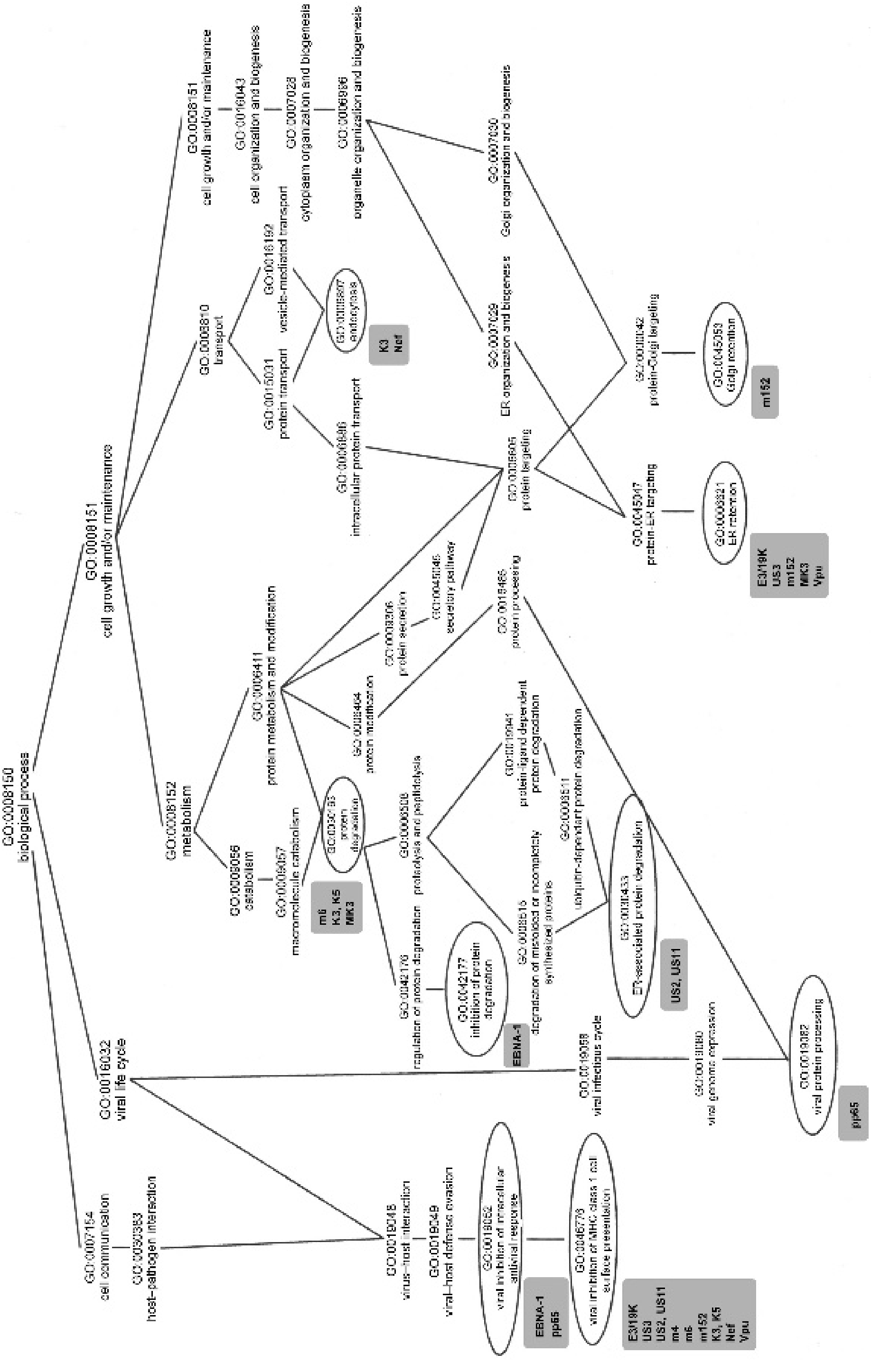

GO:0008150
biological process
GO:0007154
cell communication
GO:0030383
host–pathogen interaction
GO:0016032
viral life cycle
GO:0008151
cell growth and/or maintenance
GO:0008152
metabolism
GO:0009056
catabolism
GO:0009057
macromolecule catabolism
m6
K3, K5
MK3
GO:0030163
protein
degradation
GO:0006411
protein metabolism and modification
GO:0006464
protein modification
GO:0009306
protein secretion
GO:0045045
secretory pathway
GO:0016485
protein processing
GO:0042176
regulation of protein degradation
GO:0042177
inhibition of protein
degradation
EBNA-1
GO:0006508
proteolysis and peptidolysis
GO:0006515
degradation of misfolded or incompletely
synthesized proteins
GO:0019941
protein-ligand dependent
protein degradation
GO:0006511
ubiquitin-dependent protein degradation
GO:0030433
ER-associated protein degradation
US2, US11
GO:0019048
virus–host interaction
GO:0019049
viral–host defense evasion
GO:0019052
viral inhibition of intracellular
antiviral response
EBNA-1
pp65
GO:0046776
viral inhibition of MHC class 1 cell
surface presentation
E3/19K
US3
US2, US11
m4
m6
m152
K3, K5
Nef
Vpu
GO:0019058
viral infectious cycle
GO:0019080
viral genome expression
GO:0019082
viral protein processing
pp65
GO:0006810
transport
GO:0015031
protein transport
GO:0016192
vesicle-mediated transport
GO:0006886
intracellular protein transport
GO:0006897
endocytosis
K3
Nef
GO:0006605
protein targeting
GO:0045047
protein-ER targeting
GO:0006621
ER retention
E3/19K
US3
m152
MK3
Vpu
GO:0000042
protein-Golgi targeting
GO:0045053
Golgi retention
m152
GO:0008151
cell growth and/or maintenance
GO:0016043
cell organization and biogenesis
GO:0007028
cytoplasm organization and biogenesis
GO:0006996
organelle organization and biogenesis
GO:0007029
ER organization and biogenesis
GO:0007030
Golgi organization and biogenesis

FIG. 2. Part of the Biological Process directed acyclic graph (DAG) derived from the Gene Ontology. Each node of the DAG is labelled with the term name and GO number. From the DAG it can be seen that a parent can be linked to multiple children and different children can have multiple parents. The individual terms relating to particular viral proteins described in Table 1 are circled with the viral gene products indicated beside, in grey boxes. From the DAG the full description of the processes each viral gene targets can be obtained, for example EBNA-1 inhibits the intracellular antiviral response (GO:0019052) and inhibits protein degradation (GO:0042177). Other herpesvirus proteins target the same biological processes but at different points. This representation therefore allows cross virus functional comparisons whilst providing an updateable information resource.

TABLE 1 Representative herpesvirus, adenovirus and lentivirus proteins that interfere with antigen presentation.

Function/ Activity	*Gene/protein*	*Virus*	*Mechanism*
Effect on MHC class I	E3/19K	Adenovirus	Binding and retention of class I in ER
	US3	HCMV	Binding and retention of class I in ER
	US2, US11	HCMV	Relocation of heavy chain into ER for degradation
	m4	MCMV	Binds class I molecules
	m6	MCMV	Binding of class I molecules and transport to lysosome for degradation
	m152	MCMV	Retains class I in ER-Golgi intermediate compartment
	K3, K5	KSHV	Down-regulation of class I molecules through the ubiquitination pathway
	MK3	MHV68	Removal of class I molecules from the ER
	Nef	HIV	Endocytosis of surface class I and CD4
	Vpu	HIV	Destabilization of class I, targets CD4 to proteosome
Effect on antigen processing	EBNA-1	EBV	A Gly-Ala repeat motif prevents proteosomal degradation
	pp65	HCMV	Modulates processing of other HCMV proteins

Integrated viral data resources can provide information-rich but static data. This needs to be interpreted in the context of the dynamic process occurring in the host cells and tissues during infection and immune response. Determining which mRNAs are expressed in a cell gives an idea of which proteins are present. Large-scale gene expression mapping using gene arrays is motivated by the premise that the functional state of the organism is largely determined by its expressed genes (based on the central dogma). This means it is possible to define an organism's or cell's phenotypic state in terms of the range of genes that are expressed. This field of functional genomics has been called 'transcriptomics'.

The effects of viral infection on the transcriptome of cells *in vitro* and *in vivo* has been determined for viruses as diverse as retroviruses, herpesviruses, orthomyxoviruses, enteroviruses, adenoviruses, hepatitis B and C viruses, and papilloma viruses. These initial studies show that viruses cause both common and unique changes in cellular gene expression profiles during their replication cycle (Fruh et al 2001, Kellam 2000, 2001). However, the degree of host

transcriptional modulation is likely to be more complicated than a simple reflection of the type of virus causing the infection. Whereas viral cellular tropism is often thought of as activated by the presence or absence of an appropriate viral cell surface receptor, tropism actually includes all aspects of the cellular environment required for productive virus infection. It is possible that viruses with restricted tissue tropism reflect the reliance on a tissue or cell specific component or pathway (Sheehy et al 2002). Differential cellular gene expression is not a new concept but the extent of cellular transcriptional differences is clearly shown by microarray studies (Fig. 3). Viruses with broad tissue tropism may therefore utilize core cellular pathways present in many cell types. Alternatively viruses may be able to 'sense' a particular cellular environment and modify it towards the needs of the virus. Indications that this can happen come from studies of human cytomegalovirus (HCMV). *In vivo*, HCMV productively infects fibroblasts, epithelial cells, endothelial cells and macrophages. Uninfected fibroblasts and endothelial cells express different subsets of genes (Fig. 3a), and when infected by the same strain of HCMV elicit cell-type dependent and independent transcriptional changes (Fig. 3a). In such an example, host cell transcriptional changes can reflect either virus induced cell type specific changes to facilitate virus replication, cell type specific responses to infection, or a combination of both.

Functional genomics methods also have the potential for reunifying studies of bacteriology, virology and parasitology under an umbrella of how host and microbe interact. In this instance it is informative to examine the effect of microbial products on immune and non-immune system cell types. Again, this serves to illustrate the functional complexity of transcriptional responses to a common signal. For example, when HeLa cells (epithelial cells) and peripheral blood dendritic cells are exposed to bacterial lipopolysaccharide (LPS), cell-type specific transcriptional profiles are clearly seen (Fig. 3b). Following LPS exposure dendritic cells massively up-regulate a set of genes that are transcriptionally unresponsive in HeLa cells. In the case of dendritic cells exposed to a virus, a bacterium or a yeast, such transcriptional plasticity could be rationalized into core and pathogen-specific transcriptional responses (Huang et al 2001). There is likely to be even greater differential gene regulation *in vivo*, taking into account cell interactions, tissue location and the shifting cytokine environment. Perhaps the greatest challenge of such systems biology and immunoinformatics is to define and model such complex biology within the appropriate contextual environment of the host.

Given the inherent complexity and incomplete knowledge of host–pathogen interactions, can computational and functional genomics have an impact on therapies for otherwise difficult-to-treat viral diseases? Three examples of the strategy of using global views of differential gene expression to define drug

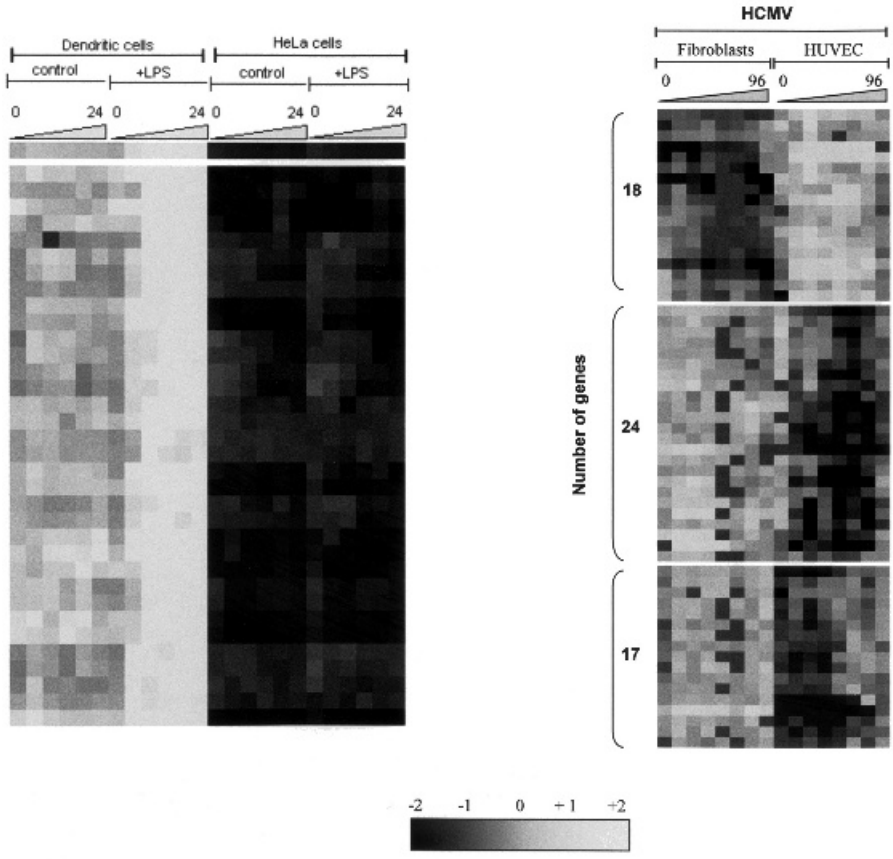


FIG. 3. Different cellular transcriptional responses to extracellular agents or virus infection measured using human gene DNA microarrays. Shades of grey indicate detectable gene expression above and black below the median expression for that gene in all the samples. (a) Human vascular endothelial cells (HUVEC) and fibroblasts display cell type specific gene expression patterns (group of 18 genes, timepoint zero). When these cells were infected with human cytomegalovirus differential gene expression occurred in the different cell types. For example in the group of 24 genes, infection of HUVECs results in a decrease in gene expression whilst the same genes remain largely unchanged in fibroblasts. Also in the group of 17 genes infection of HUVECs results in increased gene expression by 48 hours whilst there is rapid induction and then little change in the fibroblast gene expression levels. (b) Human dendritic cells and HeLa cells were mock exposed or exposed to lipopolysaccharide (LPS) over a 24 hour time course. For this group of 34 genes initially more highly expressed in dendritic cells than HeLa cells, the addition of LPS results in a massive increase in gene expression in dendritic cells but no effect in HeLa cells.

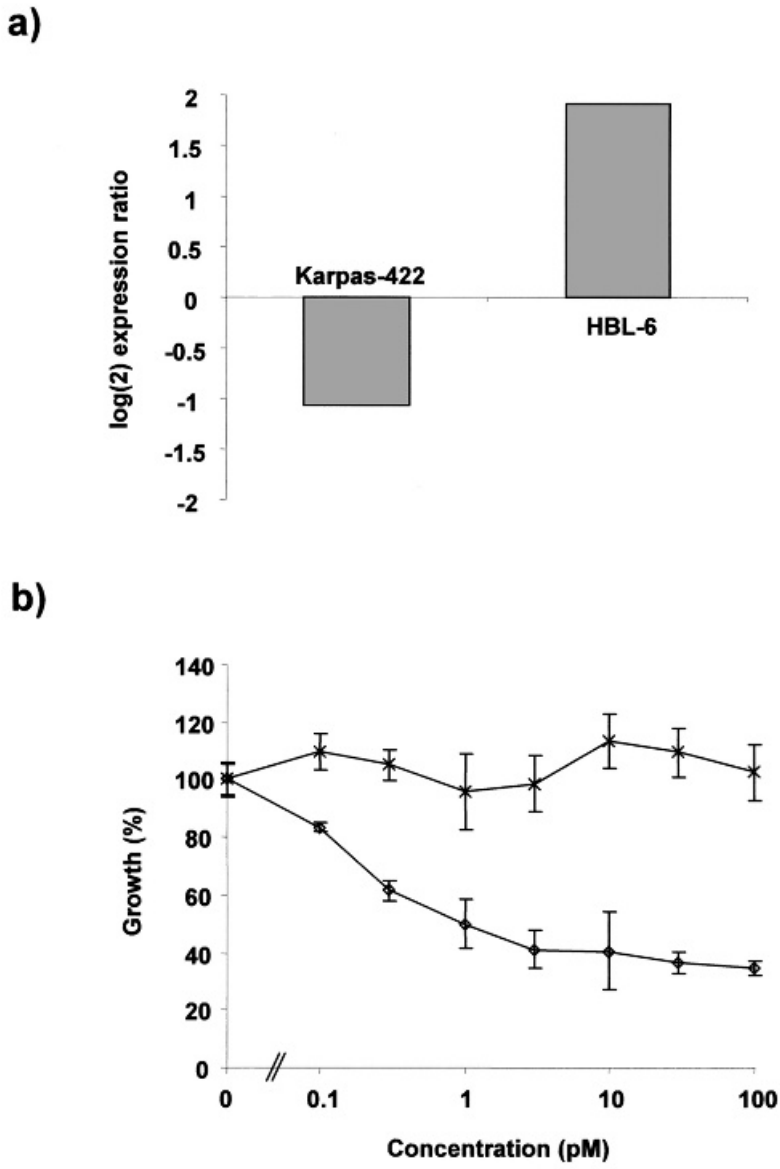


FIG. 4. Predicting drug sensitivity from gene expression pattern. (a) The relative $\log_2$ expression level of the vitamin D receptor gene in the follicular lymphoma cell line Karpas-422 and the KSHV infected primary effusion lymphoma (PEL) cell line HBL-6 showing higher expression in HBL-6. (b) Cell proliferation of HBL-6 but not Karpas-422 is inhibited in a dose dependant manner by the vitamin D analogue drug EB 1089. The proliferation assay is expressed as percentage of no drug control. (×) Karpas-422, (o) HBL-6.

targets in herpesvirus infections now exist. In the case of HCMV, infection of fibroblasts results in the induction of multiple elements of the prostaglandin E2 synthesis pathway including the gene encoding for cylo-oyxgenase 2 (COX2) (Zhu et al 1998). Inhibitors of COX2 prevent prostaglandin synthesis, and importantly COX2 inhibitors also block HCMV replication *in vitro* (Fruh et al 2001). Similarly, the proto-oncogene c-*kit* was shown to be up-regulated by KSHV during viral transformation of endothelial cells. c-*kit* is a molecular target for Gleevec which is licensed for the treatment of gastrointestinal stromal tumours (GISTs). Gleevec was shown to reverse KSHV-induced morphological transformation of endothelial cells (Moses et al 2002). In a similar study the expression profiling of viral and non-viral induced B-cell lymphomas revealed that KSHV driven lymphomas overexpress the vitamin D receptor. Targeting of these lymphomas *in vitro* with vitamin D analogue drugs prevented lymphoma proliferation in a dose-dependent manner (Fig. 4) (Jenner et al 2003). These three studies begin to offer hope that detailed viral bioinformatics and functional genomics will open up novel insights and therapeutic strategies for viral infections that are at present difficult to treat effectively.

Acknowledgments

We thank Robin Weiss for his critical review of the manuscript. This work is funded by the Medical Research Council (PK and RJ), the Biotechnology and Biological Sciences Research Council (RH), the Department of Virology, UCL (EG), the Triangle Trust and the Overseas Research Student Award (AK).

References

Albà MM, Lee D, Pearl FM et al 2001a VIDA: a virus database system for the organization of animal virus genome open reading frames. Nucleic Acids Res 29:133–136

Albà MM, Das R, Orengo CA, Kellam P 2001b Genomewide function conservation and phylogeny in the Herpesviridae. Genome Res 11:43–54

Alcami A, Koszinowski UH 2000 Viral mechanisms of immune evasion. Trends Microbiol 8:410–418

Ashburner M, Ball CA, Blake JA et al 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

Boname JM, Stevenson PG 2001 MHC class I ubiquitination by a viral PHD/LAP finger protein. Immunity 15:627–636

Coscoy L, Ganem D 2001 A viral protein that selectively downregulates ICAM–1 and B7–2 and modulates T cell costimulation. J Clin Invest 107:1599–1606

Fruh K, Simmen K, Luukkonen BG, Bell YC, Ghazal P 2001 Virogenomics: a novel approach to antiviral drug discovery. Drug Discov Today 6:621–627

Hannon GJ 2002 RNA interference. Nature 418:244–251

Holzerlandt R, Orengo CA, Kellam P, Albà MM 2002 Identification of new herpesvirus gene homologues in the human genome. Genome Res 12:1739–1748

Huang Q, Liu D, Majewski P et al 2001 The plasticity of dendritic cell responses to pathogens and their components. Science 294:870–875

Ishido S, Choi JK, Lee BS et al 2000 Inhibition of natural killer cell-mediated cytotoxicity by Kaposi's sarcoma-associated herpesvirus K5 protein. Immunity 13:365–374

Jenner RG, Boshoff C 2002 The molecular pathology of Kaposi's sarcoma associated herpesvirus. Biochim Biophys Acta 1602:1–22

Jenner RG, Maillard K, Cattini M et al 2003 Kaposi's sarcoma-associated herpesvirus-infected primary effusion lymphoma has a plasma cell gene expression profile. Proc Natl Acad Sci USA, in press

Kellam P 2000 Host–pathogen studies in the post-genomic era. Genome Biol 1:REVIEWS1009

Kellam P 2001 Post-genomic virology: the impact of bioinformatics, microarrays and proteomics on investigating host and pathogen interactions. Rev Med Virol 11:313–329

Kellam P, Albà MM 2002 Virus bioinformatics: databases and recent application. Appl Bioinform 1:37–42

Lorenzo ME, Jung JU, Ploegh HL 2002 Kaposi's sarcoma-associated herpesvirus K3 utilizes the ubiquitin–proteasome system in routing class major histocompatibility complexes to late endocytic compartments. J Virol 76:5522–5531

Means RE, Ishido S, Alvarez X, Jung JU 2002 Multiple endocytic trafficking pathways of MHC class I molecules induced by a Herpesvirus protein. EMBO J 21:1638–1649

Montague MG, Hutchison CA 3rd 2000 Gene content phylogeny of herpesviruses. Proc Natl Acad Sci USA 97:5334–5339

Moses AV, Jarvis MA, Raggo C et al 2002 Kaposi's sarcoma-associated herpesvirus-induced upregulation of the c-kit proto-oncogene, as identified by gene expression profiling, is essential for the transformation of endothelial cells. J Virol 76:8383–8399

O'Neill LA 2002 Toll-like receptor signal transduction and the tailoring of innate immunity: a role for Mal? Trends Immunol 23:296–300

Sheehy AM, Gaddis NC, Choi JD, Malim MH 2002 Isolation of a human gene that inhibits HIV–1 infection and is suppressed by the viral Vif protein. Nature 418:646–650

Thome M, Schneider P, Hofmann K et al 1997 Viral FLICE-inhibitory proteins FLIPs prevent apoptosis induced by death receptors. Nature 386:517–521

Zhu H, Cong JP, Mamtora G, Gingeras T, Shenk T 1998 Cellular gene expression altered by human cytomegalovirus: global monitoring with oligonucleotide arrays. Proc Natl Acad Sci USA 95:14470–14475

DISCUSSION

De Groot: I thought your point that you see different gene expression in different cell types was important. I had thought that things were a lot simpler. This makes me concerned about the interaction between host and pathogen, which will be different in different cell types.

Kellam: I think so. The level of complexity increases even further when you start looking at multi cell-type environments and the context of what is going on in an immune response.

De Groot: There are large viruses and small viruses. But even in HIV, our colleague Willy Hildebrand has performed an analysis of those peptides presented in the context of autologous HLA molecules in a normal cell line and those that are presented an HIV-infected cell line, and has now shown that two

different sets of epitopes derived from self proteins are processed and presented. The effect of infections on epitope processing and presentation will be complex to unravel.

Kesmir: Could this be because virally infected cells here have an immunoproteasome which cleaves differently than the constitutive proteasome? I am not sure it is entirely the effect of the virus we are seeing here.

Kellam: There are some other large-scale expression analyses that start to hint at the complexity. For example, if you take herpes simplex virus or human cytomegalovirus and expose cells to live or inactivated virus, the HCMV seems to signal through its cell surface receptors that it is attaching to and induce an interferon response inside the cell. The result is that when the virus internalizes into the cell it then disarms certain components of the cell machinery, whilst keeping certain components going. In contrast, HSV doesn't do this at the level of the cell surface, but when it is internalized it seems to induce very similar profiles to HCMV and then starts to turn off processes that it doesn't want. It has been assumed in the past that for viruses that are simple and which infect multiple cell types it is at the level of the cell surface receptor where the tropism manifests itself. But this is probably a vast oversimplification.

Petrovsky: If we are going to model immune responses, particularly to pathogens, we are going to have to be very careful that we look at the effect of the pathogens on the immune system at the same time as we are looking at the effects of the immune system on the pathogen. For example, in T cell assays people would tend to test a peptide or dead bacterium or virus, and not use live organisms, which might result in completely different responses.

Rammensee: This is being done with virus-infected cells.

Petrovsky: We tend not to put live bacteria into cultures, though.

Rammensee: No one wants bacteria in their incubators.

Wingender: We used *Pseudomonas* on cultures of epithelial cells. The results are dependent on the strain of *Pseudomonas* and also the cell type. It doesn't surprise me that we find such different responses in different cells, when we consider that one promoter exerts different effects on a gene in different cellular contexts. The same promoter is occupied by completely different sets of transcription factors in different cell lines.

Kellam: If you represent lots of knowledge, you still have to add a dynamic context in order to interpret some of the gene expression changes. There always needs to be a healthy interplay between the experimental biologist and the computational scientist.

Brusic: From the modelling perspective it seems that systems of differential equations will not take us very far. At least for some situations we need to use network models, which can describe complex interactions, including feedback loops.

Perelson: It depends on what level of information you would like to get. If you are interested in overall patterns of growth of virus and numbers of cells infected, one still might want to work at the differential equation level. If one wants to look at the responses of the host cells and the details of what is happening at the cellular level then maybe we need network models. There is room for different types of models to answer different questions.

Kellam: This raises the possibility that we can use this knowledge effectively. We can use the cellular profile in response to a pathogen as a diagnostic tool. This may allow us to determine what pathogen the immune system is seeing at the moment. There is evidence from the dendritic cells that exposure to different pathogens produces a different transcriptional profile. If you can determine such changes from the peripheral blood you might improve prognostic and diagnostic indications of different pathogens. Just by documenting what is different and what is the same we can get some immediate benefits, rather than having to go to a full network model to understand it.

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

Final general discussion

Littlejohn: I have a general observation from the perspective of a relative outsider looking in. The immune system is obviously a fantastically intricate one with a whole lot of inputs and outputs. And we are collecting and cataloguing a great many sets of data, which we are putting into a digital format so we can do immunoinformatics. In terms of thinking about the database issues I would like to see a mapping of the inputs and outputs of the immune system in a logical domain-specific classification. Then we could stipulate, for example, that we have to build 20 databases integrated in the following ways, and these are the ones we should build first. If we were to do this we could get to a point where the molecular biology informatics community is with cross-links between the major databases it uses. Is this a useful mud map to draw out?

Brusic: Tim Littlejohn is talking about a conceptual model of the immune system that is hierarchical so that we can zoom in at the specific parts at different levels. When I started working on the MHC this is exactly what I did. I drew a conceptual model of the immune system and then defined subsystem all the way to molecular interactions such as MHC–peptide binding.

Littlejohn: This could be used to derive a plan for databases.

Brusic: The advantage of this approach is that it can show inter-relationships at high levels and also at low levels. Alan Perelson and his colleagues have been working on this for ages.

Littlejohn: It is probably more of a directed acyclic graph than a hierarchy.

Perelson: I agree with you. There are a variety of levels at which we can look. As a community, people have focused in on particular molecular details, mainly emphasizing MHC and epitopes; I think there are practical issues of vaccine design driving this. There is another community that is involved with cytokines: I don't know what the status of their databases are, but they will be important to integrate. Then there is a lot of information that is just not being captured, such as functional information coming out of functional assays, which can be quite messy compared with sequence data. Investigators are now doing a lot of work measuring the functional response of T cells after they are stimulated by particular epitopes. As far as I know these data are just sitting in publications. There is also the issue of somatic hypermutation in B cells in germinal centres. There are some scattered databases gathering sequence information looking at some of the genetic changes here. It would be nice to combine this information with other pieces of the picture.

Lefranc: I think Tim Littlejohn's proposal is still a bit of a dream rather than reality, but it is nice to have this as a goal. In practice, I don't know whether it is feasible. When we try to do these things, even in a small area, very quickly we get into a situation where it is impossible to compare all the steps of the database.

Littlejohn: I would like to remind everyone of the informatics benefits here. My two favourite examples are the human mutation database initiative and the MIAME/MGED (Minimum Information About a Microarray Experiment; *http://www.mged.org/Workgroups/MIAME/miame.html*) initiative. The MIAME/MGED initiative for microarrays has derived not only minimum information capturing standards but also ontologies for the capture and storage of that information. Because of this it has the fringe benefits of XML formats and compliant databases. This is fantastic. The other good model for this community is the human mutation database, with distrubuted curation and domain experts all doing their bit. The benefits of having a group stand over it and look at the holistic problem are immense.

Schonbach: The MIAME standards are useful from a technical and implementation perspective, but we should also discuss how we can capture and compare data that are time dependent, or which are spatial data. This is much more complicated.

Gulukota: A point I would like to raise is what kind of non-immunological applications there could be for the kind of knowledge that will be derived from immunoinformatics. One example might be ways of combating computer viruses. It is not a self/non-self problem like in immunology, but it is a malicious versus non-malicious problem.

Littlejohn: That is a good thought. Think of predator–prey models like those developed for HIV battling the immune system in environmental biology. I imagine this could be a bidirectional information flow.

Flower: There are lots of computer science people interested in artificial immune systems who are using the immune system as a metaphor for doing very different things.

Perelson: There are also possibilities for using immune system analogies in pattern recognition, such as in analysing computer systems for unusual patterns that might indicate intruders.

De Groot: Another application is using tetramers for detection of acutely infected individuals. One of the benefits of this type of discussion here is that as people who are working in informatics we are pushing the limits of understanding of the immune system by asking questions that then have to be proven *invitro*. If we say that we think that most of the proteins in TB are going to be seen by the immune system, experimentalists have to go and prove us wrong. Having been here I'm starting to think that there might be an immunome in a certain state in a certain tissue, and another immunome in another state of alert in a different tissue.

Littlejohn: Another general observation is that this is the first meeting that I have been to with a domain-specific bioinformatics theme. This is tremendous. The concept of 'immunoinformatics' is very focusing. This could act as a template for other disciplines. Oncoinformatics is the next obvious one.

Wingender: I am not sure that the concepts here are not so different from the concepts arising in systems biology. In Germany a funding programme has just been launched for systems biology on the liver as an organ.

Littlejohn: The problem is that you either make it too big or too technological, and this is the intersection of a specific domain in technology.

Wingender: That is what I understood they were trying to do in these systems biology programmes as well.

Kellam: The other thing is that this is tractable. With the best will in the world you can work on the liver but you are not going to be able to do time-dependent sampling on an individual. The same is true in development: in the human system you can't do these sorts of things, quite rightly. But you would have more of an access to the immune system. It makes a good model for working in systems biology.

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

Closing remarks

Hans-Georg Rammensee

Interfakultäres Institut für Zellbiologie, Abteilung Immunologie, Universität Tübingen, Auf der Morgenstelle 15, D-72076 Tübingen, Germany

It is very difficult to sum up a meeting like this, because we are still at the beginning in this field. I think we are about where the HLA field was in 1969 when it was recognized that the subject was very complex and that it was important to give internationally recognized names to the genes and proteins, and standardize procedures for identifying these. It was also recognized that the collection of data would grow enormously. We are now in a similar position: we recognize that the procedures and names have to be standardized, and that there is a long way still to go.

In my introduction, I broke informatics down into three categories, hard, semi-soft and soft. After our discussions, I think we need to downgrade everything! The top category, hard informatics — which represents exact knowledge — probably won't ever be part of immunoinformatics. It is our aim, but it won't be reached by this field. I should add here that 'hard', 'semi-soft' and 'soft' do not mean good or bad: we need all of these categories, they are just descriptions. The top level of immunoinformatics is, I think, three-quarters hard, one-quarter soft. The semi-soft part represents the algorithms trying to predict something which is then verified experimentally, and then we have the soft category composed of the models that eventually lead to predictions. However, we now need an extra category, which I will call the 'liquid' category, a term coined by Stephan Beck: beyond the soft. This classification is more-or-less semantic, but perhaps it helps to structure our thoughts.

Something else I would like to include in this summing-up is that we have missed out quite a number of fields within immunoinformatics. Most significantly, there is the entire field of B cell immunology. Another important area not dealt with in depth is that of chemokines and cytokines. The field of signal transduction is closely linked to immunology, and I am sure that are more fields we haven't done justice to. Altogether, though, this meeting will probably inspire a lot of future activity, and I would like to thank you all for your participation.

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

Index of contributors

Non-participating co-authors are indicated by asterisks. Entries in bold indicate papers; other entries refer to discussion contributions.

Immunoinformatics: Bioinformatic Strategies for Better Understanding of Immune Function: Novartis Foundation Symposium 254. Volume 254
Edited by Gregory Bock and Jamie Goode
 ISBN: 0-470-85356-5

Subject index