Symmetry and Kinetic Optimization of Proteinlike Heteropolymers

Erik D. Nelson, Lynn F. Teneyck, and José N. Onuchic

San Diego Supercomputer Center, 10100 John Jay Hopkins Drive, La Jolla, California 92093
and Department of Physics, University of California at San Diego, La Jolla, California 92093

(Received 31 December 1996)

The relationship between optimization, evolutionary sequence selection, and structural symmetry is investigated for an elementary continuum model of proteins in which a complete correspondence between sequence and structure can be established. It is found that (i) kinetic optimization (minimal frustration) is strongly connected with ground state symmetry and (ii) the highest symmetry ground state is the least fragile to sequence mutations. [S0031-9007(97)04169-0]

PACS numbers: 87.15.By, 87.10.+e

It is generally accepted that proteins are the result of evolutionary sequence selection to optimize the stability and kinetic accessibility of the native conformation [1–18]. Nevertheless, it is clear that once a protein is able to fold and conduct its function there is no intrinsic reason why it should continue to be optimized, and therefore the space of existent proteins will, to some degree, reflect topology of sequence space in the absence of optimization pressures.

It was recently suggested [15] that structural symmetry in proteins is connected with the property of minimal frustration [6] and kinetic funneling [7], and that symmetric ground states have the largest number of sequences that will fold to them and are therefore easiest to design [19,20]. Do symmetric ground states support kinetic optimization and sequence prevalence simultaneously?

In this Letter we examine an extremely simple continuum model of a protein that allows us to introduce the competition between linear and topological memory (sequence connectivity and cross-chain contacts) as a small perturbation to the global ground state for sequences with a specific composition. The model is a continuum analog of the HP model [8] in which H and P monomers are connected by a flexible filament of fixed length. The HP interaction rule approximates the hydrophobic forces that cause proteins to collapse by giving one species of monomer (H) a much stronger mutual attraction than the other (P) which act only as obstructions. The sequence connectivity of H monomers acts as a small perturbation on the topology of the ground state unless more than two H monomers are connected sequentially in the chain. This partitions the ground states into different structures corresponding to the different sequential H segments that can appear within a sequence of specific composition.

We consider fixed composition sequences of 7 H and 9 P monomers. The collapsed states of these polymers have a core of hydrophobic monomers that at low temperatures condenses into a symmetric structure (Fig. 1) similar or identical to the ground state for 7 disconnected H monomers [21,22]. It is shown that sequences in which the pair interactions between all H monomers are equivalent (no sequential H monomers) allow for maximum symmetry, and fold into the global ground state—a pentagonal bipyramid core—while sequences that break the equivalence of H interactions (contain sequential H monomers) fold into a partially frustrated ground state with higher energy and lower structural symmetry. We study the kinetics of representative sequences from a completely classified sequence space and the evolution of sequence space in the absence of optimization pressures. It

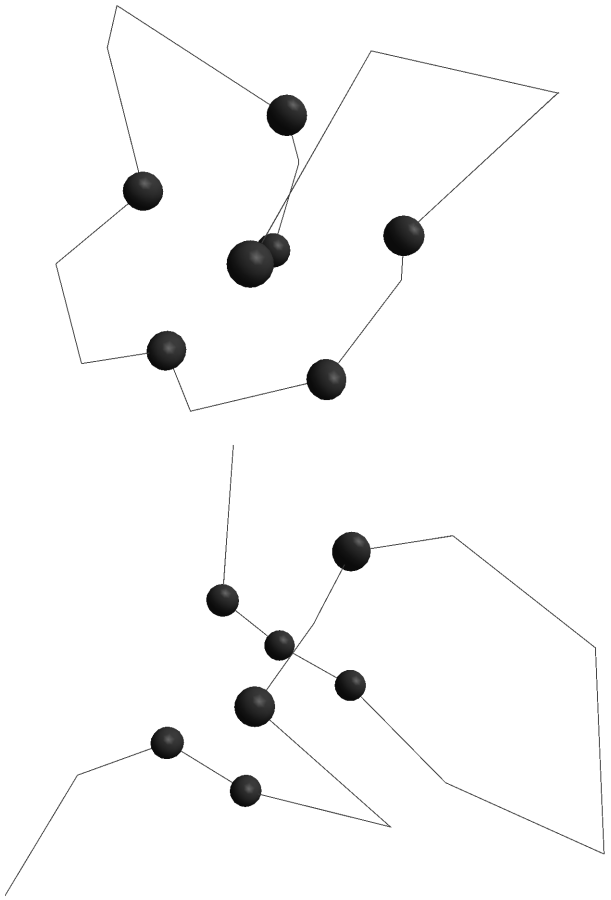

FIG. 1. Pentagonal bipyramid ground state structure $G1$ (top) and a partially frustrated ground state $G2$ (sequences $p1$ and $p2$). For clarity, P monomers have been reduced to the size of points.

 0031-9007/97/79(18)/3534(4)$10.00

is shown that (i) kinetic optimization (minimal frustration) is strongly connected with structural symmetry, and (ii) the most symmetric structure is the most robust to sequence mutations.

We begin by examining the kinetics of three representative sequences that fold to ground states with different degrees of symmetry, and then illustrate how sequence space is partitioned according to ground state structure.

The model we have constructed is a polymer chain composed of N monomers at sequence positions $1 \leq i \leq N$. Monomers that are nearest neighbors in sequence interact according to the string potential

$$\nu(x_{ij}) = 0\,, \qquad x_- < x_{ij} < x_+\,, \tag{1}$$

where x_{ij} are the nearest neighbor distances and $\nu(x_{ij} > x_+) = \nu(x_{ij} < x_-) = \infty$; in other words, the distances $x_{ij} > x_+$ and $x_{ij} < x_-$ are forbidden [23]. Monomers that are not nearest neighbors in sequence interact according to the Morse potential

$$V(r_{ij}) = \epsilon[e^{-2\lambda(r_{ij}-r_o)} - 2e^{-\lambda(r_{ij}-r_o)}], \tag{2}$$

where i and j are topological (non-nearest) neighbors and r_{ij} are their separations. The interaction parameters are constructed so that the Morse potential cores of sequence neighbors are close enough to automatically exclude self-penetration of the polymer [23] during the course of its random motion [24]. To implement the HP model we set $\epsilon = 5$ (in arbitrary units) for topological interactions so that the relevant temperatures are ~ 1, and use only the repulsive core of the potential for both HP and PP interactions.

To study the equilibrium kinetics of the polymers we recorded Monte Carlo histograms $N(E, q, T_n)$ measuring the number of times that states with potential energy E and order parameter q are visited at temperature T_n during a simulation of length t_n. For a sufficiently long simulation, $N(E, q, T_n)$ is proportional to the thermal average of the density of states $\omega(E, q)$,

$$\frac{1}{t_n} N(E, q, T_n) = \frac{1}{Z(T_n)}\, \omega(E, q) e^{-E/k_B T_n}, \tag{3}$$

where $Z(T)$ is the partition function,

$$Z(T) = \sum_{E,q} \omega(E, q) e^{-E/k_B T}. \tag{4}$$

T is the temperature, and k_B is Boltzmann's constant. From simulations at one or more T_n it is then possible to estimate $\omega(E, q)$ in order to calculate observables over a continuous range of temperatures [25,26].

The results of numerical experiments for three representative sequences with ground state structures $G1$, $G3$, and $G2$ (in descending order of symmetry) are presented below. The first sequence

$$p1 = 1\,0\,1\,0\,0\,1\,0\,0\,1\,0\,1\,0\,1\,0\,0\,1 \tag{5}$$

was constructed so that all H (1) monomers are separated by at least one P (0) monomer and, therefore, allows for HH (topological) interactions between all H monomers [27]. In this case, the H monomers behave almost like a gas of disconnected residues, and the competition between linear and topological memory [4–6] is completely minimized in the ground state $G1$ (Fig. 1).

If the obstructions of P monomers are neglected, the polymer can access a large number of energetically equivalent topological states that conform to this core structure. To enumerate these states we calculate the contact matrix: $c_{ij} = 1$ if $r_{ij} \leq a$, $c_{ij} = 0$ otherwise, for HH contacts within the $G1$ geometry (where $a \simeq 1$ is the range of the potential). Each state α corresponds to a contact matrix $c_{ij}^{(\alpha)}$ for the core and each matrix $c_{ij}^{(\alpha)}$ has 16 total contacts (note that the contact matrix is invariant with respect to sign reversal so it cannot distinguish between a conformation and its mirror image). When the geometric interference of P monomers is not taken into account, there are 252 distinct contact matrices possible for $p1$ within the $G1$ structure. Kinetically, a large number of these states are inaccessible or unfavorable so that only about 25 of them are ever occupied for a significant time.

In sequence $p1$, the competition for HH contacts is not frustrated, which results in an ensemble of maximally connected, and therefore highly symmetric ground states. However, when the equivalence of HH interactions is broken by the introduction of H-H (string) bonds, the environments of individual monomers are differentiated, and the degeneracy is suppressed.

This is the case with sequence 2,

$$p2 = 0\,1\,1\,1\,0\,1\,0\,0\,1\,0\,1\,0\,1\,1\,0\,0, \tag{6}$$

in which three nearest neighbor H-H string bonds make three of the topological interactions unavailable, frustrating the formation of the maximally connected structure. If $p2$ is constrained to the high symmetry ($G1$) core structure, each matrix $c_{ij}^{(\alpha)}$ has only 13 contacts and the degeneracy is about half as large. Frustration leads to a distorted ground state for $p2$ with lower symmetry than $G1$, but with a larger number of energetic contacts ~ 14 and much less degeneracy $= 8$.

Finally, we consider a sequence that folds to an intermediate symmetry, $G3$

$$p3 = 0\,1\,1\,1\,0\,0\,0\,0\,0\,0\,0\,0\,1\,1\,1\,1, \tag{7}$$

which is between $G1$ and $G2$, and has a unique ground state.

To calculate the occupancy of the low energy states, we computed histograms $N(E, q_\alpha, T_n)$ of the energy and the order parameter

$$q_\alpha(t) = \sum_{ij} c_{ij}(t) c_{ij}^{(\alpha)}\,, \tag{8}$$

where $c_{ij}(t)$ is the contact matrix of the polymer at Monte Carlo step t and $c_{ij}^{(\alpha)}$ is the contact matrix for

one of the degenerate conformations α of a polymer in its ground state core structure. $q_\alpha(t)$ measures the projection of sampled core conformations against the state α. The histograms were used to compute $P(\alpha, T)$, the probability that $c_{ij}(t)$ projects (completely overlaps) the state α. Figure 2 shows $P(\alpha, T)$ for $p2$ and the net probability, $P(\Gamma, T)$, that $p2$ projects any one of the states in the ground state ensemble $\Gamma = \{\alpha\}$. The folding temperature, defined as $P(\Gamma, T_f) = 1/2$, is $T_f(2) \sim 0.56$. The two most highly occupied states in the ensemble correspond to two easily interconverted conformations. The fact that the sum of these two probabilities exceeds unity below T_f reflects the fact that additional (extremely weak) contacts begin to form, resulting in a contact matrix that projects both states simultaneously.

Near T_f, the free energy $F(E, T)$ of the low symmetry sequences $p2$ and $p3$ exhibits a double well, and the distribution of energy states is bimodal. The size of the free energy barriers is related to the length of sequential H monomer segments. By contrast, the folding temperature of the high symmetry sequence $p1$ is $T_f(1) \sim 0.86$ and the free energy of $p1$ near $T_f(1)$ is smooth and downhill, in other words no free energy barrier exists between the unfolded and folded states which correspond to the Type 0 scenario proposed by Bryngelson and co-workers [13] and appears to be common for highly optimized sequences [13,14]. As one would expect, $p1$ has a lower ground state energy than both $p2$ and $p3$, and higher folding temperature than $p2$. Interestingly, the folding temperatures $T_f(1) < T_f(3) \sim 0.96$ and folding times $\tau(1) \sim \tau(3)$ of $p1$ and $p3$ are roughly the same. Although this type of fluctuation is expected to occur in short polymers [26] we also find that $G3$ is fragile to mutations. This behavior is not consistent with an evolutionarily selected folding mechanism, which should be both efficient and robust.

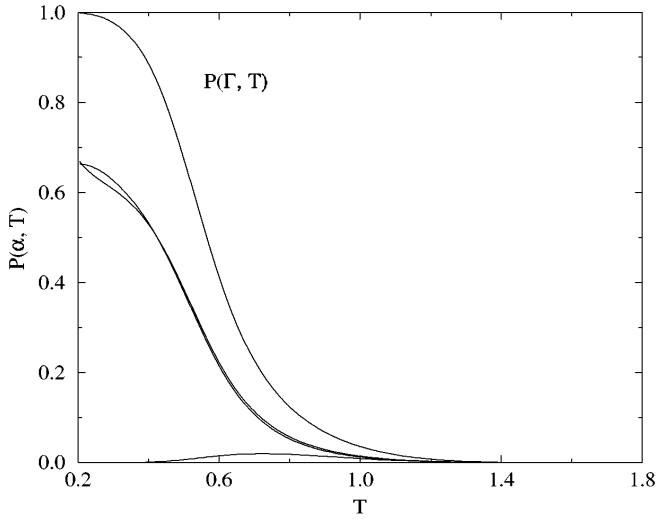

FIG. 2. Occupation probabilities $P(\alpha, T)$ and $P(\Gamma, T)$ calculated from two simulations at $T_1 = 1.15$ and $T_2 = 0.75$. $P(\alpha, T)$ measures the probability that $c_{ij}(t)$ projects (completely overlaps) the state $c_{ij}^{(\alpha)}$ while $P(\Gamma, T)$ is the probability of projecting *any* state in the ground state ensemble Γ.

In order to establish the correspondence between sequence and structure, we computed all possible 7 H monomer sequences and classified them according to the total number of H-H string bonds $s(p)$, and the number $c(p)$ of H segments (strings of 2 or more successive H monomers) in each sequence p (Fig. 3). The ground state core structure of a single representative sequence for each sequence class (15 total) was extracted as the lowest energy state observed during 10^9 Monte Carlo steps at a temperature roughly equidistant from the average collapse and folding temperatures of the sequences. It is easy to see (Fig. 1) that adding or subtracting P monomers from between any two H monomers will not affect the structure of the ground state *except* when this breaks or forms H-H bonds, but then the resulting sequence belongs in a new H segment class. Evidently, the ground state structures are partitioned only by the types of H segments present in a sequence. Thus, except for points (2, 4) and (2, 5) (which individually contain two H segment classes), each phase point corresponds to a single ground state structure.

For each sequence, we considered all possible single exchange mutations,

$$p(j) \leftrightarrow p(k) \quad \text{if } p(j) \neq p(k) \tag{9}$$

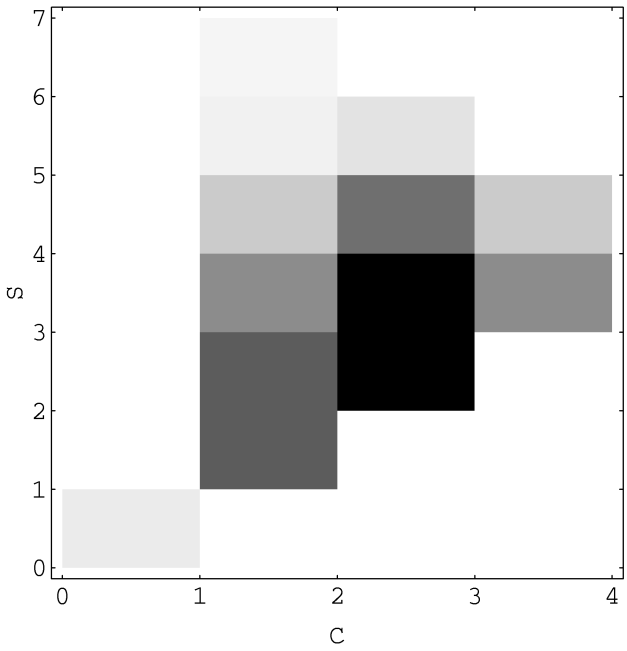

FIG. 3. The distribution of all possible sequence $N(c, s)$. For comparison, the point (2, 2) (lower black rectangle) contains 2520 different sequences while the point (1, 1) contains 1260. The total number of sequences is $\sum_{c,s} N(c, s) = 11\,440$. The points along the diagonal $c = s$ correspond to the high symmetry structure $G1$, the points (2, 3) and (3, 4) correspond to $G2$, and the points (1, 6), (1, 3), and (2, 5) correspond to $G3$ symmetry.

[where $p(j) = 0, 1$] in order to calculate the single mutation transition probability matrix

$$P(c, s \rightarrow c', s'). \tag{10}$$

From this matrix we calculate the transition probability $P(G \rightarrow G')$ between symmetry types. The diagonal elements of this matrix determine the probability of remaining within a symmetry type after a single mutation. For the three symmetries $G1$–$G3$, the probabilities are $\sim$0.7, 0.4, and 0.2, respectively. The transition probabilities $P(c, s \rightarrow c', s')$ were then used to compute the long time behavior of the resulting Markov process [28], starting from an initially uniform distribution of sequences among the phase points. For a constant mutation rate per sequence, an initially uniform distribution evolves into one similar to that shown in Fig. 3, from which it is clear that the highest symmetry structure type, $G1$, is the least fragile to mutations.

What does this model say about protein evolution? In order to answer this question we must know the degree to which evolution is driven [11] by kinetic energetic optimization, in other words, the window of conditions within which proteins are required to fold. The model above simply measures the allowed topology of sequence space, and therefore admits all possible seven H monomer structures. When we require the polymers to fold under more restrictive conditions [29] the limiting behavior of the Markov process is toward occupation of phase points (0, 0) and (1, 1), which belong to the high symmetry structure class. Thus, when we force optimization the most symmetric structure type is even more prevalent.

For this highly simplified model, we have (i) established the connection between minimal frustration and symmetry. We have shown that sequences able to maximize hydrophobic contacts are minimally frustrated and fold to a maximally symmetric ground state. We have also shown that (ii) the most symmetric ground state structure is the most robust to sequence mutations. Although these results suggest that optimization is connected with symmetry prevalence in proteins, they also indicate that symmetry prevalence can appear without any optimization beyond that needed for folding to occur.

We would like to thank Peter Wolynes, Nicholas Socci, Vijay Pande, Bob Leary, and Victor Hazelwood for stimulating discussions and helpful suggestions during the completion of this work. The simulations would have been impossible without the facilities of San Diego Supercomputer Center. Work at SDSC was supported through the NSF Grant No. BIR 9223760, and at UCSD through the NSF Grant No. MCB 9603839 and the UC/Los Alamos Research (CULAR) Initiative.

[1] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 539 (1997).
[2] C. Anfinsen, E. Haber, M. Sela, and F. H. White, Proc. Natl. Acad. Sci. U.S.A. **47**, 1309 (1961).
[3] C. Levinthal, J. Chim. Phys. **65**, 44 (1968).
[4] I. M. Lifshitz, Sov. Phys. JETP **28**, 1280 (1969).
[5] N. Go, Annu. Rev. Biophys. Bioeng. **12**, 183 (1969).
[6] J. D. Bryngleson and P. G. Wolynes, Biopolymers **30**, 177 (1990).
[7] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. U.S.A. **89**, 8721 (1992).
[8] K. F. Lau and K. Dill, Macromolecules **22**, 3986 (1989).
[9] E. Shaknovitch, G. Farztdinov, A. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).
[10] C. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. U.S.A. **90**, 6369 (1993).
[11] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Proc. Natl. Acad. Sci. U.S.A. **91**, 12 972 (1994).
[12] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Macromolecules **28**, 2218 (1995).
[13] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, Proteins **21**, 167 (1995).
[14] P. G. Wolynes, Z. Luthey-Schulten, and J. N. Onuchic, Chem. Biol. **3**, 425 (1996).
[15] P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **93**, 14 249 (1996).
[16] A. Sali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[17] M. H. Hao and H. A. Scheraga, J. Chem. Phys. **98**, 4940 (1994).
[18] E. M. Boczko and C. L. Brooks, Science **269**, 393 (1995).
[19] K. Yue and K. A. Dill, Proc. Natl. Acad. Sci. U.S.A. **92**, 146 (1995).
[20] M. E. Kellman, J. Chem. Phys. **105**, 2500 (1996).
[21] D. J. Wales, Science **271**, 925 (1996).
[22] P. A. Braier, R. S. Berry, and D. J. Wales, J. Chem. Phys. **93**, 8745 (1990).
[23] A. Milchev, W. Paul, and K. Binder, J. Chem. Phys. **99**, 4786 (1993). The Morse parameters are $\lambda = 24.0$, and $r_o = 0.8$ which produce an interaction range $a \simeq 1$ and core radius $r_c = 0.77$. The spring parameters are $x_o = 0.7$ and $x_\pm = x_o \pm 0.3$.
[24] W. P. Petersen, *Lagged Fibonacci Sequence Random Number Generators for the NEC SX3*, netlib.org (1996).
[25] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **61**, 2635 (1988); **63**, 2822 (1989).
[26] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **103**, 4732 (1995).
[27] Binary patterning of H and P residues occurs in proteins for essentially the same reason; S. Kamtekar, J. M. Schiffer, H. Ziong, J. M. Babik, and M. H. Hecht, Science **262**, 1680 (1993).
[28] C. E. Shannon, *Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949).
[29] E. D. Nelson, L. Teneyck, and J. Onuchic (to be published).