Evolution of Model Proteins on a Foldability Landscape

Sridhar Govindarajan[1] **and Richard A. Goldstein**[1,2]*
[1]*Department of Chemistry, University of Michigan, Ann Arbor, Michigan*
[2]*Biophysics Research Division, University of Michigan, Ann Arbor, Michigan*

ABSTRACT **We model the evolution of simple lattice proteins as a random walk in a fitness landscape, where the fitness represents the ability of the protein to fold. At higher selective pressure, the evolutionary trajectories are confined to neutral networks where the native structure is conserved and the dynamics are non self-averaging and nonexponential. The optimizability of the corresponding native structure has a strong effect on the size of these neutral networks and thus on the nature of the evolutionary process. Proteins 29:461–466, 1997.** © **1997 Wiley-Liss, Inc.**

Key words: protein folding; molecular evolution; lattice models; fitness landscapes; neutral networks; spin-glass theory

INTRODUCTION

Biological macromolecules are the result of eons of evolution. To understand these macromolecules, it is necessary to understand the evolutionary processes that determined their form and function. Similarly, biological evolution is constrained by the properties of the polymers that encode, represent, and manifest the evolutionary heritage; the process of evolution must be analyzed in this context.

A major step forward in our understanding of evolution was provided by the notion of a fitness landscape, representing the fitness of the system as a function of the parameters of that system.[1] Evolution then is represented by movement in that landscape. Although initial work concentrated on abstract models for the fitness landscape,[2–5] more recent models have included specific properties of the evolving macromolecules.[6–10] However, constructing an appropriate fitness function has remained elusive. One problem is that the fitness of a biological macromolecule is complicated and multifaceted. For instance, in considering protein evolution, we must simultaneously consider the ability of the protein to be synthesized, remain stable, resist proteolysis and aggregation, and fulfill the specific catalytic or structural role required of that protein.

In previous work, we developed the notion of a fitness landscape where fitness is represented by the foldability of the protein as a function of the parameters defining the intramolecular interactions between residues.[11–13] Although it is only one aspect of the variety of properties required by any protein, the ability to fold to a stable native state is universal and nontrivial and is a major factor separating biological proteins from random sequences of amino acids. In addition, the fact that proteins of similar structures often fulfill different functions and similar functions are sometimes performed by proteins of radically different structure indicates that the structural properties of evolved proteins are largely determined by structural, rather than functional, requirements. This model represents an approximation that the majority of the protein structure serves as a scaffold to maintain the active site in a correct conformation, and thus the selective pressure acting on functionality is restricted to a small percentage of the residues in the protein.

By concentrating on foldability, we are able to take advantage of the development of simple models that have greatly increased our understanding of the folding process. Furthermore, by adjusting the degree of foldability considered adequate, we can explore how the evolutionary process was affected by the degree of pressure acting on selection.

By considering the range of interactions corresponding to different native structures, we were able provide a reason why certain structural motifs are so commonly found among biological proteins.[11,12] We also showed how this simple model results in biological proteins having marginal stability, that is, the minimum thermodynamic stability necessary for the protein to be adequately foldable and stable.[13] Most interestingly, we found that higher degrees of selective pressure corresponds to evolutionary trajectories that are confined to “neutral networks,” paths through the interaction space where changes in sequence and interaction parameters do not result in changes in the native structure.[13] This effect can

Contract grant sponsors: College of Literature, Science, and the Arts, the Program in Protein Structure and Design, the Horace H. Rackham School of Graduate Studies; Contract grant sponsor: NIH; Contract grant number: LM0S77; Contract grant sponsor: NSF; Contract grant number: BIR9512955.

*Correspondence to: Richard A. Goldstein, Department of Chemistry, University of Michigan, Ann Arbor, MI 48109-1055. E-mail: richardg@Umich.edu

Received 24 April 1997; Accepted 2 July 1997

explain the presence of proteins with common native structures yet with dissimilar sequences and stabilizing interactions.

In this study, we focus our attention on trajectories of random walks in this landscape and note the presence of glassy behavior at higher degrees of selective pressure, evident from the smaller rate of change of both the interactions and the structures and from the nonexponential and non self-averaging behavior of the evolution. We explore the nature of the neutral networks and show how the size of these neutral paths for different native structures depends critically on the maximum foldability possible for that structure. This observation provides an explanation based on evolutionary dynamics of why certain structures are overrepresented among biological proteins and suggests that there should be large differences in the evolutionary behavior of proteins with different structures.

METHODS

Our study is based on lattice models of proteins, where the possible conformations of a 27-residue protein are represented by the 103,346 self-avoiding walks on a 3 × 3 × 3 lattice, with each residue occupying a single lattice site. Although the noncompact states are important for understanding the folding process, these states will have less effect on the qualitative nature of the foldability landscape. On the basis of earlier work, we assume that the energy of the protein is dominated by pair contacts,[14] which depend only on the identity of the two residues; the energy function for any sequence in conformation m is then given by

$$E = \sum_{i<j} \gamma(A_i, A_j)\Delta_{ij}^{m} \qquad (1)$$

where $\gamma(A_i, A_j)$ is the contact potential between residue type A_i at position i in the sequence and residue type A_j at position j, and $\Delta_{ij}^{m} = 1$ if residues i and j are not adjacent in sequence but are on adjacent lattice sites in conformation m, and zero otherwise. Every compact state has exactly 28 formed contacts. We use the parameter values derived by Miyazawa and Jernigan[15] for $\gamma(A_i, A_j)$, which implicitly include the effect of interactions with the solvent. We can calculate the energy of any sequence k in all possible compact states by using the above energy function. The native structure N^k is assumed to be the conformation of lowest energy.

We are interested in characterizing the ability of a protein to fold, an explicitly kinetic property. Although simulations of evolution based on selective pressure to maximize the folding rate have been described,[16] it is currently unfeasible to perform the characterizations that we are interested in by modeling the dynamic characteristics of the protein at each step in the evolutionary trajectory. Instead, we take advantage of the extensive computational and theoretical analyses concerning the relationship between the thermodynamic properties of a protein and the ability of that protein to fold. This work was originally pioneered by Bryngelson and Wolynes,[17] based on a theoretical treatment using concepts borrowed from the physics of spin glasses. In their study, they showed that the ability of a protein to fold into its correct native state and to avoid the local minima (modeled as a transition to a glassy state) is dependent on the relative ratio of the folding transition temperature to the glass transition temperature. Wolynes and co-workers[18,19] then showed that this ratio of temperatures could be maximized by maximizing the ratio of the energy gap between the native state and the average of the nonnative states to the standard deviation in the energies of these nonnative conformations.[18,19] Monte Carlo simulations of the folding of lattice proteins have shown that a large value of this latter ratio distinguished proteins that could fold from those that could not.[20–24] We characterize each sequence by a "foldability" F^k representing this ratio. The minimum value of this parameter required for sufficiently rapid folding is termed the critical foldability F_{crit}. We easily can simulate different degrees of selective pressure by modifying the value of the critical foldability.

The distance measure between two native structures N^k and N^l is given by the total number of contacts common to the two native structures divided by 28, the total number of contacts in all of the compact states:

$$q_{kl} = \frac{1}{28} \sum_{i<j} \Delta_{ij}^{N^k} \Delta_{ij}^{N^l}. \qquad (2)$$

Identical structures have a q value of 1. The natural distance metric for biological sequences k and l is the Hamming distance h_{kl}, representing the number of mutations necessary to change one sequence into the other. Unfortunately, this metric has limitations when dealing with protein sequences. Some amino acids are quite similar (threonine and serine), whereas others are radically different (cysteine and phenylalanine), so mutations can be either conservative or nonconservative. In addition, the impact of any mutation depends on the other residues in the protein. For this reason, as in previous work, we consider a separate interaction space, representing the set of values of the pairwise contact potentials $\{\gamma_{ij}\} = \{\gamma(A_i, A_j)\}$ for every pair of residues i and j.[11–13] Because contacts can be formed only between a residue in an even position and a nonadjacent residue in an odd position, there are exactly 156 possible contacts that can be formed, meaning that the intramolecular interactions for any sequence can be specified by the 156 values of γ_{ij}. The possible values of these interaction parameters thus defines a 156-

dimensional space, with specific sequences representing discrete points in this space. Because each compact state has exactly the same number of contacts and the foldability involves a ratio of energies, the native state and the foldability are not affected by scaling all of the γ_{ij} values with either an additive or multiplicative constant. For this reason, we normalize the interaction parameters so that $\Sigma_{i<j}\,\gamma_{ij} = 0$ and $\Sigma_{i<j}\,\gamma_{ij}^2 = 1$, effectively projecting the points in the interaction space to the surface of a unit hypersphere. We then can measure the distances between points in the interaction space corresponding to sequences k and l by considering the angular distance θ_{kl} between these two points on this hypersphere. Due to the high dimensionality of this space, the distances between pairs of sequences are strongly clustered around $\theta_{kl} = \pi/2$.[13]

The fitness landscape can be characterized by the behavior of walks on that landscape. We start with an initial random amino acid sequence and observe what happens as that sequence is mutated. We consider only single site mutations, where the rest of the sequence is held fixed. The mutation results in a new sequence, with a new foldability $F\ '$ and possibly a new native structure $N\ '$. The attempted mutation is only accepted if the new foldability is higher than the critical foldability F_{crit}. A "generation" is counted for every attempted mutation, whether or not it is accepted. For each different value of F_{crit}, simulations of 10,000 generations were performed for five different initial sequences selected at random. A simple hill-climbing scheme was used to generate an initial sequence with a sufficiently large value of F_{crit}. One hundred generations were simulated to allow the system to equilibrate before statistics were accumulated.

RESULTS

For values of F_{crit} less than approximately 6.0, $\langle\cos(\theta)\rangle_t$, the average cosine of the distance in interaction space between points separated by t generations, is well represented as an exponentially decaying function of t with a diffusion constant D_θ. For values of F_{crit} above 6.0, $\langle\cos(\theta)\rangle_t$ becomes better approximated by a stretched exponential of the form

$$\langle\cos(\theta)\rangle_t = e^{-D_\theta t^\beta} \tag{3}$$

where β decreases to a value of 0.37 at $F_{\text{crit}} = 7.0$. We define a generalized diffusion constant D_θ^g as

$$D_\theta^g = \left[\int_0^\infty e^{-D_\theta t^\beta}\,dt\right]^{-1} \tag{4}$$

which reduces to the standard diffusion constant for exponential decays. A similar generalized diffusion constant D_q^g is defined for changes in the native conformation, with

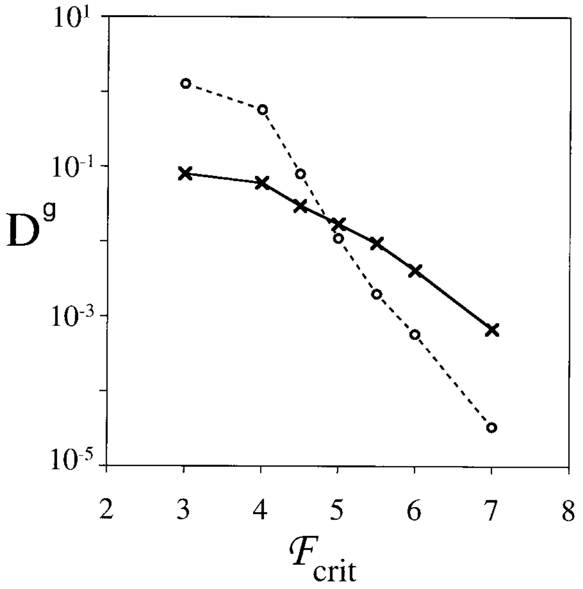

Fig. 1. Generalized diffusion constants for diffusion in interaction space, D_θ^g (—) and diffusion in structure space, D_q^g, (- - -) as a function of F_{crit}. As shown, diffusion in structure space is more dependent on F_{crit} than is diffusion in interaction space.

$$\langle q\rangle_t = (1 - \overline{q})e^{-D_q t^\beta} + \overline{q} \tag{5}$$

where $\langle q\rangle_t$ is the average q value between two structures separated by t generations and $\overline{q}$ is the average q value between two random structures. D_θ^g and D_q^g are plotted on a logarithm scale in Figure 1 as a function of F_{crit}. As shown, the slopes of the two curves are significantly different, D_q^g showing a much greater dependence on the selective pressure than D_θ^g. The consequence of this is that under conditions of large selective pressure, the structures are "frozen in," whereas appreciable changes in the sequence and intramolecular interactions still occur. This result indicates that confinement to a single structure during evolution may be a consequence of the requirement that the protein must be able to fold to a unique conformation, and not necessarily due to structural constraints on the protein.

The nonexponential diffusion in interaction space at high selective pressure is characteristic of situations where the dynamics start to be dominated by the roughness of the fitness landscape. Another characteristic is the presence of non self-averaging dynamics; that is, the ensemble average no longer equals the time average. Both the nonexponential evolutionary dynamics and this non self-averaging behavior is shown in Figure 2, which shows $\langle\cos(\theta)\rangle_t$ for different individual runs at two different values of F_{crit}. In addition to the slower movement in the interaction space, the dynamics are highly depen-

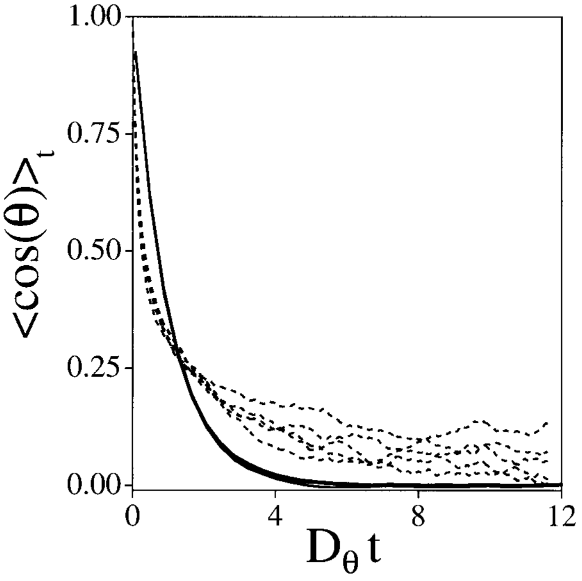

Fig. 2. $\langle\cos(\theta)\rangle_t$, the average cosine of the distance between sequences in interaction space separated by t generations, scaled by the generalized diffusion constant, for five different runs at two values of F_{crit}: $F_{\mathrm{crit}} = 4$ (—) and $F_{\mathrm{crit}} = 6$ (- - -). For larger values of the selective pressure, the dynamics are highly nonexponential and nonself-averaging.

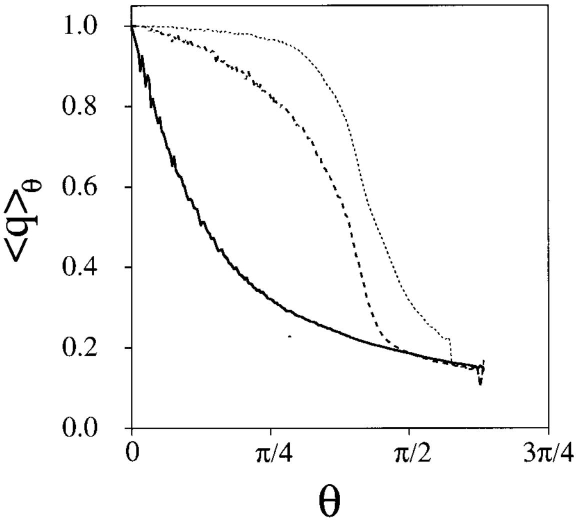

Fig. 3. Average similarity of native structures $\langle q\rangle_\theta$ for sequences separated by a distance θ in interaction space, during simulations performed with three different values of F_{crit}. When $F_{\mathrm{crit}} = 3$ (—), memory of the initial structure is quickly lost. For $F_{\mathrm{crit}} = 5$ (- - -) and $F_{\mathrm{crit}} = 6$ (· · ·), diffusion is increasingly dominated by "neutral paths," which allow changes in sequence and intramolecular interactions but retain memory of the initial structure.

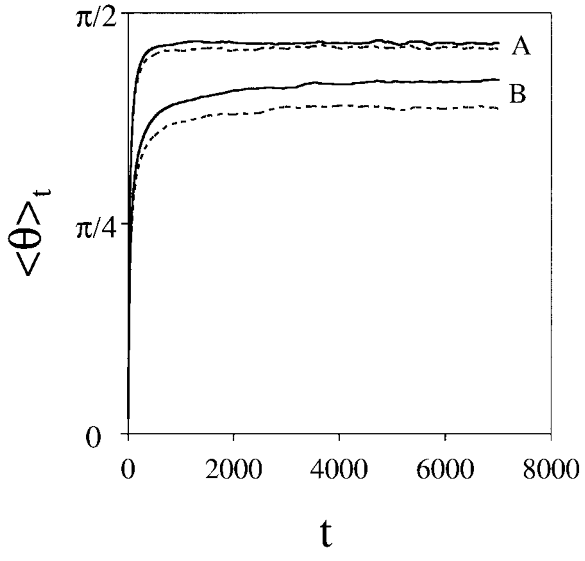

Fig. 4. Average distance in interaction space $\langle\theta\rangle_t$ for sequences separated by t generations of neutral evolution, at two different F_{crit} values; $F_{\mathrm{crit}} = 3$ (**A**) and $F_{\mathrm{crit}} = 6$ (**B**), for two different structures, one with a high F_{opt} value (—) and the other with a low F_{opt} value (- - -). These simulations were performed by only accepting mutations that preserved adequate foldability as well as native structure. At higher values of F_{crit}, the size of the neutral network depends strongly on F_{opt}.

dent on the initial starting sequence for high values of F_{crit}.

As shown in Figure 3, for low values of F_{crit}, movement in the interaction space quickly causes changes in native structure. Under conditions of higher selective pressure, corresponding to a larger value of F_{crit}, the mutational paths still cover large ranges of interaction space, yet the structures tend to become highly conserved. At an F_{crit} value of 6.0, the interactions between different sequences in the simulation can differ by sizable amounts before memory of the initial structure is lost. The larger value of F_{crit} confines the evolutionary path to "neutral networks," paths through the fitness landscape where structure is preserved, even with large changes in sequence and interaction parameters.

We can categorize the various native structures by considering the set of interaction parameters $\{\gamma_{ij}\}_{\mathrm{opt}}$ that maximize the foldability F for that structure.[11,13] Assuming a Gaussian distribution of energies of the random states, the set of optimal parameters can be solved for in closed form.[18,19] $F^{\ k}_{\mathrm{opt}}$ is the maximum foldability of native state $N^{\ k}$ when the interactions are optimal. Because the interaction space is continuous, although actual sequences exist only as discrete points in this space, the optimal set of parameters likely will not correspond to any actual sequence. Despite this fact, we have emphasized the importance of the maximum foldability as an important parameter characterizing different structural motifs.[11,12] Specifically, we find that the size of the neutral network depends on the value of $F^{\ k}_{\mathrm{opt}}$. This observation was made by performed simulations where site mutations were accepted only if the resulting fitness was larger than F_{crit} *and* the structure remained constant. Figure 4 shows $\langle\theta\rangle_t$, the average value of θ for sequences separated by t

generations for two native structures with different values of F_{opt} for two different values of F_{crit}. When F_{crit} is small ($F_{\mathrm{crit}} = 3.0$), the range in interaction parameters for both native structures can reach as far as $\langle\theta\rangle_t = 0.46\pi$, showing that the networks are extensive and are spread throughout the interaction space. When the selective pressure is high, corresponding to a large value of $F_{\mathrm{crit}} = 6$, the range of interaction space for the native structure with the higher F_{opt} value is substantially larger than for the native structure with the lower F_{opt} value. Due to the high dimensions of the interaction space, small differences in the value of θ corresponds to large differences in the volume in the interaction and sequence spaces; the networks corresponding to higher F_{opt} structures encompass much larger volume of both.

CONCLUSION

As shown, under conditions of high selective pressure, corresponding to a large value of F_{crit}, evolutionary trajectories are confined largely to neutral networks where sequence and interactions change without a corresponding change in structure. This effect can explain how the robustness of protein structures can coexist with the observed plasticity of sequences in biological proteins, a phenomenon referred to as "structural inertia."[25] These results are also consistent with the observation that different proteins with similar structures are stabilized by quite different interactions.[26] The confinement of the evolutionary trajectories to these neutral networks can have a large impact on the evolutionary dynamics, as emphasized by a number of investigators.[6,9,27,28] For instance, the large size of the neutral networks possible make neutral mutations more likely, supporting the neutral drift theory proposed by Kimura[29] and King and Jukes.[30] Additionally, because different structures would have vastly different sizes of neutral networks, one would observe considerable differences in the evolutionary dynamics and hence in the quantities, such as substitution rates and genotypical variation among various proteins. It is important to note that the size of the neutral networks in interaction space as examined here is perhaps more important than the corresponding networks in sequence space; conservative mutations between nearly identical amino acids that do not result in changed interactions will have few consequences regarding the evolution of structure and function.

It has been recognized that certain structural motifs are overrepresented among biological proteins. In earlier work, we sought to explain this effect by considering the volume of interaction space corresponding to various native structures.[11,12] We postulated that this volume would be a strong function of F_{opt}; as a result, it is exactly those structures with large values of F_{opt} that would be overrepresented. This relied on a static model of molecular evolution and ignored the dynamics, which is the essence of the evolutionary process. Studies of the dynamics, as shown here, provide qualitatively similar results based on the concept of neutral networks. For larger values of F_{crit}, the size of the neutral network explored during evolution depends critically on the F_{opt} of the corresponding structure; highly optimizable structures with large values of F_{opt} will have correspondingly larger neutral networks, resulting in a larger range of sequences that would successfully fold into that structure. This makes such structures more likely to result from evolution. Similarly, the larger neutral networks will make these proteins more robust to evolutionary changes. Additionally, highly optimizable structures with larger sized neutral networks are more likely to have a greater number of "neighboring" structures that can be reached through a valid evolutionary trajectory. These more connected structures will have greater opportunity to result from mutations of other viable structures, again suggesting that these structures might be overrepresented in the database. Finally, as mentioned above, proteins have to fulfill many requirements besides folding, including fulfilling a prescribed function. The larger neutral networks corresponding to foldability means that these structure will have much greater flexibility in adapting to these other needs.

ACKNOWLEDGMENTS

We thank Kurt Hillig and James Raines for computational assistance, and Erich Bornberg-Bauer and Jeffery Koshi for helpful discussions. Financial support was provided by the College of Literature, Science, and the Arts, the Program in Protein Structure and Design, the Horace H. Rackham School of Graduate Studies, NIH grants LM0577, and NSF equipment grant BIR9512955.

REFERENCES

1. Wright, S. The roles of mutation, inbreeding, crossbreeding and selection in evolution. Int. Proc. Sixth International Congress Genetics 1:356–366, 1932.
2. Macken, C.A., Perelson, A.S. Protein evolution on rugged landscapes. Proc. Natl. Acad. Sci. USA 86:6191–6195, 1989.
3. Derrida, B., Peliti, L. Evolution in a flat fitness landscape. Bull. Math. Biol. 53:355–382, 1991.
4. Bak, P., Flyvbjerg, H., Lautrup, B. Coevolution in a rugged fitness landscape. Physiol. Rev. A 46:6724–6730, 1992.
5. Kauffman, S.A. "The Origins of Order." New York: Oxford University Press, 1993.
6. Lipman, D.J., Wilbur, W.J. Modelling neutral and selective evolution of protein folding. Proc R Soc Lond [Biol] 245:7–11, 1991.
7. Fontana, W., Stadler, P.F., Tarazona, P., Weinberger, E.D., Schuster, P. RNA folding and combinatory landscape. Physiol. Rev. E 47:2083–2099, 1993.
8. Fontana, W., Konings, D.A.M., Stadler, P.F., Schuster, P. Statistics of RNA secondary structures. Biopolymers 33: 1389–1404, 1993.
9. Schuster, P., Fontana, W., Stadler, P.F., Hofacker, I.L. From sequences to shapes and back: A case study in RNA second-

ary structures. Proc. R. Soc. Lond. [Biol.] 255:279–284, 1994.

10. Renner, A., Bornberg-Bauer, E. Exploring the fitness landscapes of lattice proteins. In: “Pacific Symposium on Biocomputing ’97.” Altman, R.B., Dunker, A.K., Hunter, L., Klein, T.E. (eds.). Singapore: World Scientific 1996:361–372.
11. Govindarajan, S., Goldstein, R.A. Searching for foldable protein structures using optimized energy functions. Biopolymers 36:43–51, 1995.
12. Govindarajan, S., Goldstein, R.A. Why are some protein structures so common? Proc. Natl. Acad. Sci. USA 93:3341–3345, 1996.
13. Govindarajan, S., Goldstein, R.A. The foldability landscape of model proteins. Biopolymers 42:427–438, 1997.
14. Govindarajan, S., Goldstein, R.A. Optimal local propensities for model proteins. Proteins 22:413–418, 1995.
15. Miyazawa, S., Jernigan, R.L. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. Macromolecules 18:534–552, 1985.
16. Gutin, A.M., Abkevich, V.I., Shakhnovich, E.I. Evolution-like selection of fast-folding model proteins. Proc. Natl. Acad. Sci. USA 92:1282–1286, 1995.
17. Bryngelson, J.D., Wolynes, P.G. A simple statistical field theory of heteropolymer collapse with application to protein folding. Biopolymers 30:171–188, 1990.
18. Goldstein, R.A., Luthey-Schulten, Z.A., Wolynes, P.G. Optimal protein folding codes from spin glass theory. Proc. Natl. Acad. Sci. USA 89:4918–4922, 1992.
19. Goldstein, R.A., Luthey-Schulten, Z.A., Wolynes, P.G. Protein tertiary structure recognition using optimized hamiltonians with local interactions. Proc. Natl. Acad. Sci. USA 89:9029–9033, 1992.
20. Fukugita, M., Lancaster, D., Mitchard, M.G. Kinematics and thermodynamics of a folding heteropolymer. Proc. Natl. Acad. Sci. USA 90:6365–6368, 1993.
21. Šali, A., Shakhnovich, E.I., Karplus, M.J. Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. J. Mol. Biol. 235:1614–1636, 1994.
22. Šali, A., Shakhnovich, E.I., Karplus, M.J. How does a protein fold. Nature 369:248–251, 1994.
23. Chan, H.S., Dill, K.A. Transition states and folding dynamics of proteins and heteropolymers. J. Chem. Phys. 100: 9238–9257, 1994.
24. Betancourt, M.R., Onuchic, J.N. Kinetics of proteinlike models: The energy landscape factors that determine folding. J. Chem. Phys. 103:773–787, 1995.
25. Aronson, H.E.G., Royer, W.E., Jr., Hendrickson, W.A. Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. Protein Sci. 3:1706–1711, 1994.
26. Laurents, D.V., Subbiah, S., Levitt, M. Different protein sequences can give rise to highly similar folds through different stabilizing interactions. Protein Sci. 3:1938–1944, 1994.
27. Schuster, P., Stadler, P.F. Landscapes: Complex optimization problems and biopolymer structures. Comput. Chem. 3:295–324, 1994.
28. Hunyen, M.A., Stadler, P.F., Fontana, W. Smoothness within ruggedness: The role of neutrality in adaptation. Proc. Natl. Acad. Sci. USA 93:397–401, 1996.
29. Kimura, M. Evolutionary rate at the molecular level. Nature 217:624–626, 1968.
30. King, J.L., Jukes, T.H. Non-Darwinian evolution. Science 164:788–798, 1969.