Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 3341–3345, April 1996
Biophysics

Why are some protein structures so common?

(tertiary structure/protein evolution/lattice models/fitness landscapes/spin glasses)

SRIDHAR GOVINDARAJAN* AND RICHARD A. GOLDSTEIN*†

*Department of Chemistry, †Biophysics Research Division, University of Michigan, Ann Arbor, MI 48109-1055

Communicated by Peter G. Wolynes, University of Illinois, Urbana, IL, December 22, 1995 (received for review August 9, 1995)

ABSTRACT **Many biological proteins are observed to fold into one of a limited number of structural motifs. By considering the requirements imposed on proteins by their need to fold rapidly, and the ease with which such requirements can be fulfilled as a function of the native structure, we can explain why certain structures are repeatedly observed among proteins with negligible sequence similarity. This work has implications for the understanding of protein sequence–structure relationships as well as protein evolution.**

Understanding the relationship between protein sequences and structures has been a major goal of modern molecular biophysics. One of the most intriguing aspects of this problem is that the wide range of possible biological sequences (1) fold into such a small number of native structures (2–4). While arguments have been presented why certain motifs are more likely than others (5–7), the extreme degeneracy of the mapping of sequence to structure has not been explained.

There are many examples of proteins of similar structure with completely different functions and similar functions that are performed by different proteins with different tertiary folds (8). This suggests that the dominant explanation for the limited number of folds is based on structural rather than functional grounds. One of the basic requirements of any protein is that it must find its native structure in a biologically relevant time scale. Recent work with simple theoretical models has demonstrated that rapid folding to a consistent final shape can be achieved as long as this native state is sufficiently stable relative to the ensemble of random conformations (9–18). There has been some preliminary work indicating that this stabilization may be easier to achieve for some structures than for others (5–7, 19, 20).

We have been investigating how a protein sequence's ability to fold depends upon the interactions between the residues. We use concepts borrowed from the physics of spin glasses to model this foldability. According to this picture, there are two transitions possible for the sequence: (*i*) to a folded state at temperature T_f and (*ii*) to a glassy misfolded state at temperature T_g. It is the ratio of these two temperatures that determines whether the protein will be able to fold sufficiently rapidly (10, 12, 21, 22). By using the random energy model (23), T_f/T_g can be expressed as a monotonically increasing function of $R = \Delta/\Gamma$, where Δ is the difference between the energy of the native structure and the average energy of the ensemble of random conformations, and Γ is the width of the distribution of energy values of the random structures (12, 20–22). The foldability can be characterized by the value of R.

Recent work has looked at the relationship between RNA sequence and secondary structure (24). This has been facilitated by methods that can quickly find the lowest-energy structure by analyzing possible base pairings (25, 26). Such an approach is difficult with protein tertiary structure due to the number of possible conformations for a protein of even modest length. We have addressed this issue by analyzing a lattice model of proteins confined to a $3 \times 3 \times 3$ cubic lattice, where each residue is confined to a lattice point and only the interactions between the nonbonded near neighbors are considered (20). Such models can represent larger proteins where each lattice point corresponds to the position of a structural unit stabilized by local cooperative interactions, such as parts of α-helices (27). The advantage of this model is that we can do an exhaustive enumeration of all compact conformations and, if the native state and the random states are compact, we can solve for the optimal set of interaction parameters that maximizes the foldability R (12, 21, 22). We found that some structures are more optimizable than others and conjectured that there should be a connection between how much a given structure could be optimized for folding and how likely that structure would result from random evolution. This was based on the idea that for highly optimizable proteins, the interactions could be far from optimal and still result in rapid folding; structures with lower optimal folding abilities would have to have close to optimal interactions to fold.

We now develop a simple model to quantify this conjecture, by using our lattice model to characterize foldability. We find that even in the absence of any evolutionary pressure, it is still the most optimizable structures that are likely to result from evolution. This dependence grows stronger as the evolutionary pressure increases, resulting in highly optimizable structures being greatly overrepresented in any set of biological proteins. In particular, we show that while in our model all conformations are possible, certain conformations would be so overrepresented that relatively few distinct protein structural motifs would be observed in any random set of biological proteins. This can then reconcile the plasticity of protein sequences with the robustness of observed structures.

MODEL AND METHODS

We consider the space of all possible sequences of amino acids, or more abstractly, the d-dimensional space of all possible sets of interactions between residues ($\{\gamma\}$) that can characterize such sequences. Each possible protein sequence corresponds to some point in this interaction space. There will be certain regions of this space where the corresponding protein sequences would fold into some structure, separated by regions where the protein would not be sufficiently stable or where folding would occur too slowly. The volume of the space of acceptable interactions for folding into some protein structures will be larger than for others. We make the simple approximation that the probability of any given structure resulting from evolution is proportional to the volume of the interaction space that would result in successful folding into that structure.

For each of the M possible structures, we consider a separate foldability function $R_i(\{\gamma\})$ representing the ability of a protein with interactions $\{\gamma\}$ to fold into structure i. For our lattice model, the foldability would be the ratio of Δ/Γ; other models of protein folding would yield different foldability functions. If the sequence can fold, it will fold into the structure with the

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

maximum foldability for the particular set of interactions $\{\gamma\}$; that is, it will fold into structure i if $R_i(\{\gamma\}) > R_j(\{\gamma\})$ $(\forall j \neq i)$. Evolutionary pressures are represented by a constraint on the foldability necessary for the sequence to fold into a stable native structure—the protein can fold only if $R_i(\{\gamma\})$ is greater than some critical value R_c. R_c depends on the length and overall flexibility of the protein sequences and on the rate of processes such as aggregation and degradation that compete with folding. Larger values of R_c represents the situation when it is more difficult to design a foldable protein sequence, and more of the interaction space would not correspond to any foldable structure.

We characterize each of the structures by its optimal foldability R_i^o, defined as the foldability of structure i for the optimal set of interactions for that structure $(\{\gamma\}_i^o)$. The values of the optimal foldabilities are distributed according to some random probability distribution $P(R^o)$. We are interested in $\Omega_{\mathcal{F}_i}(R_i^o)$, the average volume of the interaction space that would result in successful folding into structure i $(\mathcal{F}_i)$ as a function of R_i^o. This is equal to

$$\Omega_{\mathcal{F}i}(R_i^o) = \int_{R_c}^{R_i^o} \Omega(R_i|R_i^o) P(N_{R>Ri} = 0) dR_i, \qquad [1]$$

where $\Omega(R_i|R_i^o)$ is the volume of space with foldability R_i around a point with optimal foldability R_i^o, and $P(N_{R>R_i} = 0)$ is the probability that the number of structures with higher foldability $(N_{R>R_i})$ is equal to zero. This expression assumes that the region of relevant interaction space is large enough so that structure i is competing with many other structures, so correlations in $P(N_{R>R_i} = 0)$ can be neglected. As structures with foldability less than R_c will not fold, and R_i cannot exceed R_i^o, this integral is from R_c to R_i^o.

$\Omega(R_i|R_i^o)\, dR_i$ is the volume of the d-dimensional hyperspherical shell with thickness dR_i, where the radius of the shell is such that the foldability in the shell is R_i. Assuming a simple Gaussian shape for the foldability optima,

$$R_i(\{\gamma\}) = R_i^o \exp\left(-\frac{|\{\gamma\} - \{\gamma_i^o|^2}{2\Lambda^2}\right), \qquad [2]$$

where $|\{\gamma\} - \{\gamma\}_i^o|$ is the distance between $\{\gamma\}$ and $\{\gamma\}_i^o$ in the d-dimensional interaction space and Λ is a constant parameter,

$$\Omega(R_i|R_i^o)\, dR_i = \begin{cases} \dfrac{(2\pi)^{\frac{d}{2}}\Lambda^d}{\left(\frac{d}{2}-1\right)!R_i}\left[\ln\left(\dfrac{R_i^o}{R_i}\right)\right]^{\frac{d-2}{2}} dR_i & | \quad R_i < R_i^o \\ 0 & | \quad R_i > R_i^o \end{cases}. \qquad [3]$$

We assume that the various optima are uncorrelated, so that $N_{R>R_i}$ can be described by a Poisson distribution: $P(N_{R>R_i} = 0) = e^{-\langle N_{R>R_i}\rangle}$, where $\langle N_{R>R_i}\rangle$ is the average number of structures with $R > R_i$. For the simplest approximation, we imagine that various structures have their optimal interactions randomly spread throughout the landscape with some density of optima ρ. The average number of structures at any point in interaction space whose foldability is between R and $R + dR$, and whose optimal foldability lies between R° and $R^\circ + dR^\circ$ is equal to the volume of interaction space surrounding the point of interest where an optima of optimal foldability R° in that region would result in a foldability equal to R at the point of interest [equal to $\Omega(R|R^\circ)$] times the probability of any optima having maximum foldability R° $[P(R^\circ)]$ times the overall density of optima (ρ). Integrating this expression over all values of R and R° where $R' < R < R^\circ$ yields

$$\langle N_{R>R'}\rangle = \rho \int_{R'}^{\infty} dR^o \int_{R'}^{R^o} dR\, \Omega(R|R^o)\, P(R^o). \qquad [4]$$

The distribution of $P(R^\circ)$ values can take a variety of functional forms (28). The distribution of optimal foldability values computed by using spin-glass theory based on lattice models of proteins follows a roughly Gaussian distribution (20). By using this functional form, we write:

$$P(R^o) = \frac{1}{\sqrt{2\pi}\lambda} exp\left(-\frac{(R^o - \overline{R^o})^2}{2\lambda^2}\right). \qquad [5]$$

As mentioned above, we assume that $P(\mathcal{F}|R^\circ)$, the conditional probability of folding into a particular structure given its optimal foldability R° is proportional to the volume of interaction space that would result in that particular structure; that is, $P(\mathcal{F}|R^\circ) = K\Omega_{\mathcal{F}}(R^\circ)$, where the normalization constant K can be computed by noting that the *a priori* probability of forming any particular structure is simply the reciprocal of the total number of structures M. $P_{\mathcal{F}}(R^\circ)$, the probability of a protein folding into some structure with optimal foldability R°, is just the probability of any structure having optimal foldability R° times the conditional probability of folding into that structure given that optimal foldability, summed over the M possible structures:

$$P_{\mathcal{F}}(R^o) = MP(\mathcal{F}|R^o)P(R^o). \qquad [6]$$

We can obtain $\mu(l)$, the average number of different structures that are observed l times when we pick ν points from the volume of interaction parameter space that would result in successful folding (corresponding to ν proteins with solved structures) by computing the probability of observing any structure with a given R° value l times summed over the M structures with their corresponding distribution of optimal foldabilities. Again, assuming Poisson statistics:

$$\mu(l) = \frac{M}{l!}\int_0^{\infty} [\nu P(\mathcal{F}|R^o)]^l e^{-\nu P(\mathcal{F}|R^o)}\, P(R^o)\, dR^o. \qquad [7]$$

The total number of different structures observed is:

$$\mu(l > 0) = M\int_0^{\infty} [1 - e^{\nu P(\mathcal{F}|R^o)}] P(R^o) dR^o. \qquad [8]$$

RESULTS

The various quantities described in the previous section were obtained numerically by using the parameters from the lattice model referred to above (20). In this model the variable parameters $(\{\gamma\})$ are the 156 possible interactions between residues. Since scaling the interactions or adding a constant does not change the value of R, we set the average value of the interaction parameters equal to zero and scale the parameters so that $\Sigma_k \gamma_k^2 = 1$. The interaction space is then the surface of a 155-dimensional unit hypersphere ($d = 154$). The curvature of the interaction space was neglected. The resulting distribution of optimized interaction parameters for the 103,346 possible compact structures is shown in Fig. 1. The distribution of R° values had a mean of $\overline{R^\circ} = 12.44$ with standard deviation $\lambda = 0.37$. R_c was set equal to 5.44 to give a value of T_f/T_g of 1.67 (12, 20) in agreement with estimates for real proteins (27). The average width of the optima (Λ) was estimated to be approximately 0.67, based on the percentage of structures with random interaction values that had R values greater than 5.44.

The distribution of R° values for the native protein structures $[P_{\mathcal{F}}(R^\circ)]$ for various values of the critical foldability R_c is shown

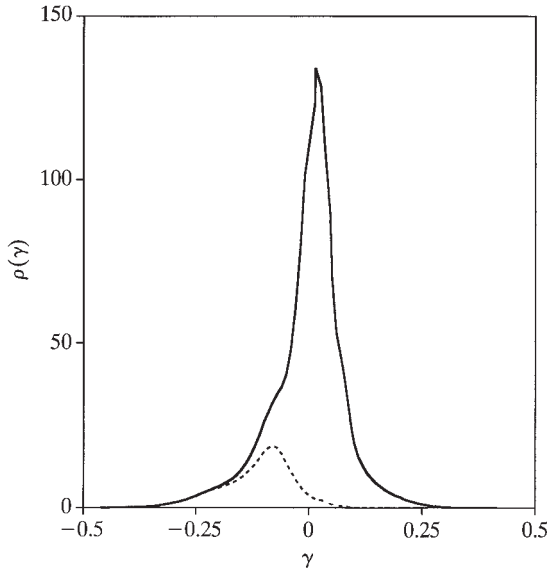

FIG. 1. Distribution of optimized interaction parameters (γ) for the 103,346 compact structures on the 3 × 3 × 3 cubic lattice, optimized by using spin-glass theory (12, 20–22) and normalized to the number of interactions in the model. The distribution for all contacts (solid line) is compared with the distribution of interaction parameters for the native contacts found in the lowest-energy structure (dashed line).

in Fig. 2. Even with no evolutionary pressure ($R_c = 0$), highly optimizable structures are more likely to result from evolution than poorly optimizable structures are. This effect becomes more extreme with increasing R_c. These results are qualitatively similar to numerical results shown in Fig. 3: 150,000 sets of interaction parameters were chosen at random for the lattice model proteins, and the $R°$ value for the resulting lowest-energy structure was calculated, by assuming no evolutionary pressure. This effect increased as the evolutionary pressure was increased (results not shown).

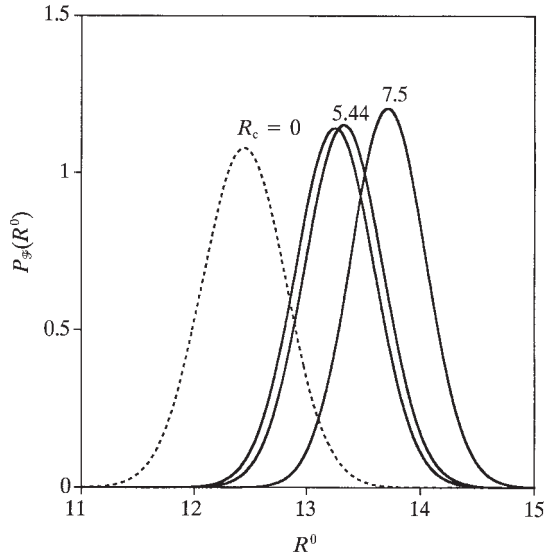

FIG. 2. Distribution of $R°$ values for the native state structures [$P_{\mathcal{F}}(R°)$] expected for various values of R_c (solid lines), compared with the distribution of $R°$ values for all possible structures [$P(R°)$] (dashed line). Even with no evolutionary pressure ($R_c = 0$), it is still the most optimizable structures that are most likely to result with a random set of interaction parameters. As the evolutionary pressure increases (increasing R_c), the distribution of the native states moves toward the extreme tail of the $P(R°)$ distribution, resulting in fewer different structures observed.

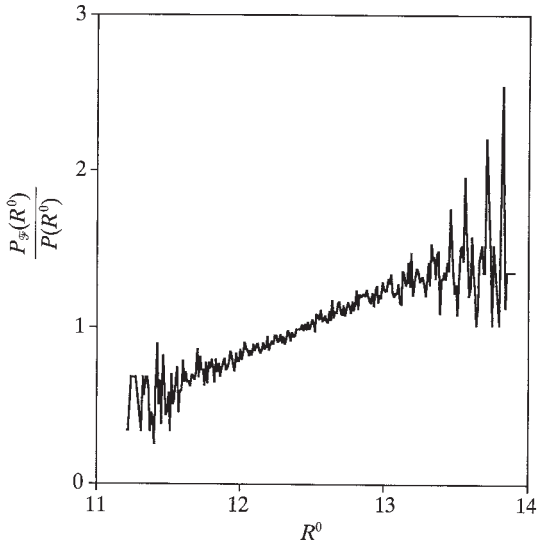

FIG. 3. Probability of the native state structure having a given $R°$ value [$P_{\mathcal{F}}(R°)$] compared with what would be expected at random [$P(R°)$], obtained by choosing 150,000 random sets of interaction parameters for the lattice model as described (20). As expected, structures that are more optimizable with larger values of $R°$ are overrepresented in the set of native structures.

As R_c increases, the corresponding shift in $P_{\mathcal{F}}(R°)$ to the tail of the $P(R°)$ distribution results in fewer different observed structural motifs and thus more examples of each. Fig. 4 shows the number of different motifs that would be observed as the number of solved structures increases, for various values of R_c. A R_c value of 5.44 would result in 1000 unrelated proteins of solved structures having a total of 789 different folds. Results are also shown for a model assuming 2025 possible equally likely structural motifs, also resulting in 789 different structures for 1000 solved proteins. Fig. 5 shows how certain motifs start to be greatly overrepresented, even for moderate R_c values, relative to what would be expected with a smaller number of equally likely structures.

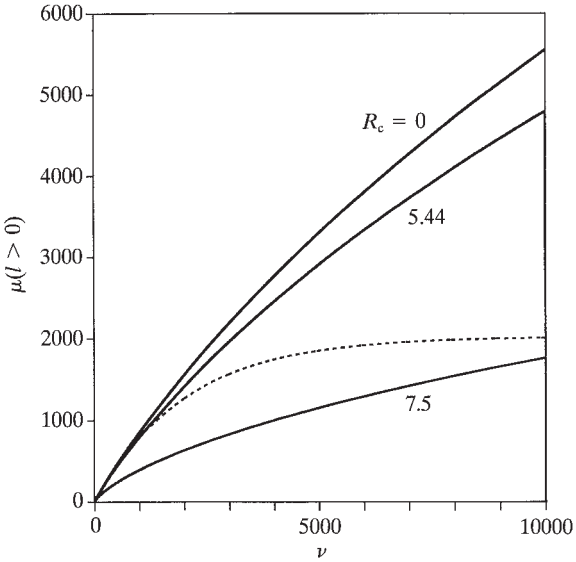

FIG. 4. Number of different structural motifs observed [$\mu(l > 0)$] as a function of the number of individual unrelated proteins of known structure (ν), for various values of R_c (solid lines). Also shown for comparison is the predicted number of different structural motifs observed if each of 2025 structures were equally likely (dashed line), a model that would mean that the observation of new structural motifs would become increasingly rare.

DISCUSSION

In our model, it is the structures that are highly optimizable that are more likely to arise through random evolution. We have used parameters obtained from our highly simplified lattice model to quantify these predictions. In contrast to many models that emphasize the role of local interactions (29–32), our model suggests that it is those conformations with many long-range contacts that would be the most optimizable, and the local propensities would be a rather small contribution to the overall stabilization of the native state (33). These conclusions have been supported by recent dynamics simulations (34). It is also interesting to note that many three-dimensional lattice structures that are highly optimizable, such as Greek key and jelly-roll motifs, are also rather common among biological proteins (20).

In contrast to models that explain the limited number of observed structural motifs by postulating a limited number of possible motifs (7), in our model, essentially all structures are possible, given the right set of interactions. It is the uneven probabilities of finding any particular structure that results in the observation of relatively few folds. As a result, as more protein structures are solved, novel motifs will continue to be observed. This model would also predict that a few folds would be observed very frequently, while most other folds would be observed only once in unrelated proteins. This seems to be a common phenomenon in the data base of known protein structures, as the work of Orengo *et al.* (4) indicates.

This work involves numerous simplifications about the nature of protein evolution. The first approximation, that the volume of interaction space resulting in a foldable protein of a given structure represents the probability that such a structure would arise through molecular evolution, represents the long-time limit where evolution has occurred to a sufficient extent that a rough degree of equilibrium over the interaction space has been reached. The approximation that the optimal sets of interactions for different structures are distributed randomly through the interaction space is supported by the observation that the pair-correlation function for the optimal set of interactions for the 103,346 structures on the $3 \times 3 \times 3$ lattice is almost identical to that obtained by selecting points at random from the corresponding hypersurface. It is interesting to note that there is a small tail to the distribution, corresponding to similar structures that are closer to each other in interaction space than would be expected at random. The role of these correlations may be important in understanding the dynamics of molecular evolution, just as the "funnels" in the free-energy landscape seem to possibly play a role in the dynamics of folding of individual proteins (35, 36).

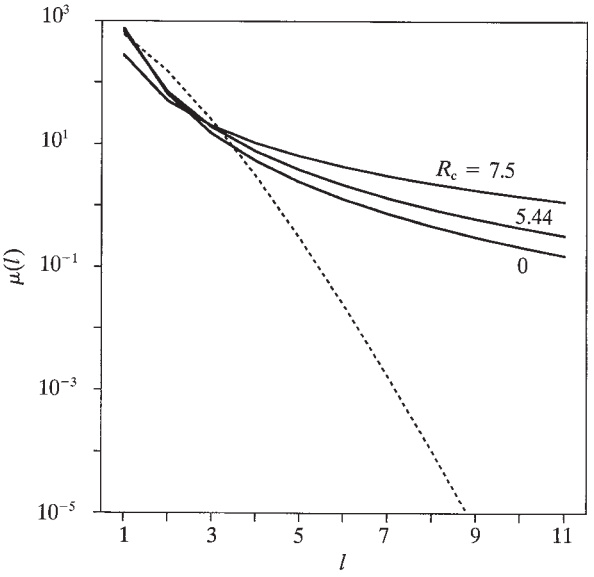

FIG. 5. Number of structures observed l times [$\mu(l)$], for 1000 solved protein structures, for various values of R_c (solid lines). Shown for comparison is the distribution when each of 2025 structural motifs is equally likely (dashed line), resulting in a much larger percentage of motifs observed multiple times, with little probability of folds observed a large number of times ($l > 6$). Such models cannot explain the existence of "superfolds" as observed by Orengo *et al.* (4).

We have also neglected the role of other forms of evolutionary pressure, such as the need for the folded protein to fulfill its biological function. Being able to fold is crucial but is only one of a long list of requirements a protein must fulfill. Our assumption is that the requirements of foldability are uncorrelated with these other requirements—two different structures are *a priori* equally likely to be compatible with a given biological activity, so the predominance of one structure over another represents how easy it is to find a set of interactions that will fold into that structure. Again, this is likely an oversimplification but is supported by the observation that different motifs have been found to fulfill the same function (8). In addition, other requirements such as biochemical efficacy could be added to the fitness function used in this model.

In the highly dimensional interaction space, almost all of the volume of the interaction space corresponding to foldable proteins would have R values only slightly greater than R_c, suggesting that proteins would be marginally foldable. In the spin-glass model, where foldability reflects how well a protein is able to avoid spin-glassy behavior, this would suggest that such behavior might be incipient. Proteins have been shown to exhibit such spin-glass behavior, especially at low temperatures (37, 38). Wolynes and coworkers (27, 36) in particular have emphasized the role of the glass transition in understanding the final stages of protein folding.

Percolation theory has been applied to the understanding of molecular evolution (39) and can provide insight into the dependence of protein evolution on the value of R_c. For low values of R_c, most of the interaction space would be accessible, and it would be rather easy for proteins to evolve from one structure to another. As R_c increased, the interaction space would move toward isolated regions representing foldable proteins separated by larger regions of interaction space corresponding to nonfoldable proteins. It is possible that nature has worked to optimize R_c through evolution so that R_c is close to the percolation threshold, where regions of interaction space corresponding to different structures are isolated enough to give a robustness to structural forms, yet not so isolated that evolution of the structure is impossible.

We thank Jeffrey Koshi and Michael Thompson for helpful comments and Kurt Hillig for computational assistance. Financial support was provided by the College of Literature, Science, and the Arts and the Program in Protein Structure and Design at the University of Michigan, and by National Institutes of Health Grant 1R29 LM05770-01.

1. Davidson, A. R. & Sauer, R. T. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 2146–2150.
2. Levitt, M. & Chothia, C. (1976) *Nature (London)* **261,** 552–557.
3. Chothia, C. (1992) *Nature (London)* **357,** 543–544.
4. Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994) *Nature (London)* **372,** 631–634.
5. Finkelstein, A. V. & Ptitsyn, O. B. (1987) *Prog. Biophys. Mol. Biol.* **50,** 171–190.
6. Finkelstein, A. V., Gutin, A. M. & Badretdinov, A. Y. (1993) *FEBS Lett.* **325,** 23–28.
7. Finkelstein, A. V., Gutin, A. M. & Badretdinov, A. Y. (1995) *Subcell Biochem.* **24,** 1–26.
8. Branden, C. & Tooze, J. (1991) *Introduction to Protein Structure* (Garland, New York).
9. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7524–7528.

10. Bryngelson, J. D. & Wolynes, P. G. (1990) *Biopolymers* **30,** 171–188.
11. Honeycutt, J. D. & Thirumalai, D. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 3526–3529.
12. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 4918–4922.
13. Fukugita, M., Lancaster, D. & Mitchard, M. G. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 6365–6368.
14. Shakhnovich, E. I. (1994) *Phys. Rev. Lett.* **72,** 3907–3910.
15. Sali, A., Shakhnovich, E. I. & Karplus, M. J. (1994) *J. Mol. Biol.* **235,** 1614–1636.
16. Sali, A., Shakhnovich, E. I. & Karplus, M. J. (1994) *Nature (London)* **369,** 248–251.
17. Chan, H. S. & Dill, K. A. (1994) *J. Chem. Phys.* **100,** 9238–9257.
18. Betancourt, M. R. & Onuchic, J. N. (1995) *J. Chem. Phys.* **103,** 773–787.
19. Yue, K. & Dill, K. A. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 4163–4167.
20. Govindarajan, S. & Goldstein, R. A. (1995) *Biopolymers* **36,** 43–51.
21. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 9029–9033.
22. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, Peter G. (1993) *Proceedings of the 26th Annual Hawaii International Conference on System Sciences*, eds. Mudge, T. N., Milutinovic, V. & Hunter, L. (IEEE Computer Soc. Press, Los Alamitos, CA), Vol. 1, pp. 699–707.
23. Derrida, B. (1980) *Phys. Rev. Lett.* **45,** 79–82.
24. Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. (1994) *Proc. R. Soc. London B* **255,** 279–284.
25. Zuker, M. & Stiegler, P. (1981) *Nucleic Acids Res.* **9,** 133–148.
26. Zuker, M. & Sankoff, D. (1984) *Bull. Math. Biol.* **46,** 591–621.
27. Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z. & Socci, N. D. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 3626–3630.
28. Gumbel, E. J. (1958) *Statistics of Extremes* (Columbia Univ. Press, New York).
29. Wetlaufer, D. B. (1973) *Proc. Natl. Acad. Sci. USA* **70,** 697–701.
30. Zwanzig, R., Szabo, A. & Bagchi, B. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 20–22.
31. Dill, K. A., Fiebig, K. M. & Chan, H. S. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 1942–1946.
32. Srinivasan, R. & Rose, G. D. (1995) *Proteins* **22,** 81–99.
33. Govindarajan, S. & Goldstein, R. A. (1995) *Proteins* **22,** 413–418.
34. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1995) *J. Mol. Biol.* **252,** 460–471.
35. Leopold, P. E., Montal, M. & Onuchic, J. N. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 8721–8725.
36. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995) *Proteins* **21,** 167–195.
37. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. (1991) *Science* **254,** 1598–1603.
38. Subramaniam, V., Bergenhem, N. C. H., Gafni, A. & Steel, D. G. (1995) *Biochemistry* **34,** 1133–1136.
39. Kauffman, S. A. (1993) *The Origins of Order* (Oxford Univ. Press, New York).