Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 9206–9209, September 1995
Biophysics

Lattice model for rapidly folding protein-like heteropolymers

INDIRA SHRIVASTAVA*, SARASWATHI VISHVESHWARA*, MAREK CIEPLAK†, AMOS MARITAN‡, AND JAYANTH R. BANAVAR§

*Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India; †Institute of Physics, Polish Academy of Sciences, 02-668 Warsaw, Poland; ‡Dipartimento di Fisica, International School for Advanced Studies, Trieste, Italy; and §Department of Physics and Center for Materials Physics, 104 Davey Laboratory, The Pennsylvania State University, University Park, PA 16802

Communicated by Gordon A. Baym, University of Illinois, Urbana, IL, June 22, 1995

ABSTRACT **Protein folding is a relatively fast process considering the astronomical number of conformations in which a protein could find itself. Within the framework of a lattice model, we show that one can design rapidly folding sequences by assigning the strongest attractive couplings to the contacts present in a target native state. Our protein design can be extended to situations with both attractive and repulsive contacts. Frustration is minimized by ensuring that all the native contacts are again strongly attractive. Strikingly, this ensures the inevitability of folding and accelerates the folding process by an order of magnitude. The evolutionary implications of our findings are discussed.**

Understanding the mechanism of protein folding is a fundamental problem in molecular biology (1). Key issues include the prediction of three-dimensional protein structure from one-dimensional sequence information and the inverse problem of designing sequences that fold into a desired three-dimensional structure. One of the puzzles of the dynamics of protein folding, called the Levinthal paradox (2), is the rapidity with which a polypeptide chain folds into the native state even though an exhaustive search is ruled out, due to the enormous number of all possible conformations. More generally, the dynamics of protein folding is one of a class of optimization problems involving the minimization of a fitness function characterized by complex structure and many local minima (for example, see refs. 3–8).

It is well known that lattice models, in spite of their simplicity, capture many of the features of protein folding (9–12). Sali, Shakhnovich, and Karplus (SSK) (13, 14) have carried out extensive studies of a simple model of a 27-bead self-avoiding chain on a simple cubic lattice to determine what the thermodynamic and kinetic requirements are to obtain rapid folding to the global energy minimum. They find that as long as the native state is a pronounced global minimum on the potential surface and the temperature is high enough to overcome the barriers between local minima, the ground state is reached within 50 million time steps in a Monte Carlo (MC) simulation—the protein is said to have folded. The SSK study (13, 14) used overall attractive contact energies (denoted by B_{ij}) similar to those of real proteins as described by Miyazawa and Jernigan (15) that were randomly assigned to the monomers of the self-avoiding chain and captured the hydrophobic effect in globular proteins. The attractive interactions lead to compact conformations of the polypeptide chain being favored and allow for a complete enumeration of the low-energy portion of the conformation space.

A protein contains 20–25% of acidic and basic groups (16) that are ionized or protonated under physiological conditions. The repulsion between like charges leads to the interactions becoming frustrated. [Iori *et al.* (17) have shown that linear heteropolymers with quenched frustrated interactions have a unique folded phase.] In addition, frustration in heteropolymers can arise from geometrical effects of excluded volume or an inability to satisfy all the hydrophobic and hydrophilic interactions simultaneously. Wolynes and collaborators (8, 18–21) have shown that the principle of minimum frustration distinguishes between natural proteins and random heteropolymers—in proteins, side chains contribute coherently to supersecondary structures and there is a harmonious relation between secondary structures and tertiary folds (9). This principle leads naturally to a large stability gap that is a measure of the energy gap between the states with a structure similar to that of the native state and the lowest energy state among those that bear little structural resemblance to the native structure. In simple situations, the stability gap may be correlated with the energy gap (13, 14) between the native and the first excited state. Goldstein *et al.* (20, 21) have used the maximization of the stability gap of proteins of known sequences and structures as a means of determining the optimal interactions between amino acids and have successfully predicted folding structures for new sequences.

A key issue is whether this propensity of proteins to fold rapidly is due to the evolutionary selection of sequences that tend to have larger stability gaps than random sequences and how one may design such proteins. We will demonstrate that an idea similar to the notion of strong disorder in spin glasses (22, 23) can indeed be used to design protein sequences that are strongly folding. When the exchange interactions in an Ising spin glass are widely separated, frustration, or the inability of a spin to satisfy all the exchange interactions of its neighbors simultaneously, while present, is irrelevant and the nontrivial ground state of the spin glass can be obtained trivially. Operationally, this can be achieved by rank-ordering the exchange interactions in decreasing magnitude and arranging the mutual spin orientations to satisfy as many of these as possible in order of decreasing strength. We will use this rank-ordering idea in our protein design—however, we will not require the monomer interactions to be widely separated. We first select a compact conformation as the target structure of the folded protein and identify the contacts—two monomers i and j are in contact with a contact energy B_{ij} if they are not successive in sequence and yet next to each other. We work with the same distribution of B_{ij} values as SSK (13, 14). There are 28 contacts for compact self-avoiding chain conformations. Our 27-monomer chain on a simple cubic lattice has 14 odd-numbered monomers and 13 even-numbered monomers. The two-sublattice structure of the lattice ensures that odd-numbered monomers have only even-numbered monomers as neighbors and vice versa. Thus the total number of possible contacts is 156 obtained by noting that each of the 13 even-numbered monomers can possibly have at most 12 contacts (monomers that are successive in sequence are not considered contacts). We rank order the 156 B_{ij} values in decreasing order (by magnitude) and assign the first 28 values randomly to 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.

Abbreviations: SSK, Sali, Shakhnovich, and Karplus; MC, Monte Carlo.

Table 1. Summary of the results

Sequence	E_0	Foldicity	CC	Gap	$\mathrm{Min_{MC}}$	$\mathrm{Max_{MC}}$	$T(X = 0.8)$
			Purely attractive				
1	−95.606	1.0	0.615	7.385	0.66	21.81	2.77
2	−90.282	0.9	0.576	2.623	1.00	26.70	2.30
3	−92.613	0.8	0.610	5.267	0.14	5.38	2.83
4	−93.614	1.0	0.622	6.282	0.10	6.23	2.92
5	−93.220	1.0	0.598	8.058	0.54	32.89	2.66
6	−100.843	0.9	0.638	6.859	0.59	23.73	3.25
7	−91.879	0.9	0.574	4.318	0.57	8.57	2.50
8	−96.076	1.0	0.595	6.043	0.46	13.58	2.82
9	−91.702	1.0	0.587	3.794	0.07	4.01	2.53
10	−98.524	1.0	0.575	8.983	0.05	3.01	3.39
			Partially repulsive				
1	−76.686	1.0	0.597	11.013	0.0058	0.401	6.37
2	−81.672	1.0	0.592	18.235	0.0299	0.731	6.12
3	−82.069	1.0	0.590	15.338	0.0818	0.162	5.78
4	−89.214	1.0	0.590	13.653	0.0746	1.228	6.53
5	−87.411	1.0	0.605	8.686	0.0422	1.126	6.47
6	−84.543	1.0	0.627	13.165	0.0292	6.262	6.19
7	−82.726	1.0	0.596	11.501	0.0454	0.890	6.90
8	−79.187	1.0	0.599	14.288	0.0110	0.086	6.47
9	−79.289	1.0	0.586	15.489	0.0304	0.168	6.04
10	−84.221	1.0	0.580	16.136	0.0263	0.292	5.73

MC simulations were carried out as described in refs. 13 and 14. The MC procedure starts with a random self-avoiding conformation. Local MC moves of one or two successive monomers that maintain bond length and the self avoidance of the chain are carried out with the Metropolis algorithm. For each sequence, 156 random B_{ij} values with a Gaussian distribution of mean −2 and width 1 were generated. In the partially repulsive case, half of these numbers were reversed in sign. E_0 denotes the native state energy. A sequence is defined to have folded if it reaches the native state within 50 million MC steps. The foldicity of a sequence is defined as the fraction of 10 MC runs starting from different initial configurations that succeeds in folding. CC denotes the Pearson correlation coefficient (24) between the interaction matrix and the native contact map. The gap denotes the energy difference between the native state and the first excited state among the compact self-avoiding conformations. $\mathrm{Min_{MC}}$ and $\mathrm{Max_{MC}}$ are in units of 10^6 MC steps and denote the minimum and maximum time required to fold the sequence from 10 starting conformations. (Only the cases in which folding took place within 50 million MC steps have been considered.) The order parameter (7) is $X(T) = 1 - \Sigma p_i^2$, where $p_i = \exp(-E_i/k_BT)/Z$, $Z = \Sigma\exp(-E_i/k_BT)$, E_i is the energy, and only maximally compact self-avoiding conformations are considered. $X(T)$ is a measure of the degeneracy of the heteropolymer chain and may be used to characterize the transition between a state with many equivalent conformations to a unique folded state. We find a small value of X over a wide range of low temperatures indicating the relative stability of the native state. The last entry in the table denotes the temperature at which $X = 0.8$. The value of $T(X = 0.8)$ is significantly higher for the partially repulsive case underscoring the relative stability of the native state among the compact conformations. The MC simulations were carried out at a temperature of 1.3, which is the mean value of $T(X = 0.8)$ for the folding sequences in the SSK study (13, 14). It is important to note that, in our studies, this temperature is low enough that the native state is more stable than the denatured states (26, 27). Considering only the maximally compact conformations, the probability that the system is in the native state is equal to 1/2 at temperatures between 2.1 and 6.7 for the cases studied. The energy gaps between the native state and the first excited state among the maximally compact conformations are larger than in the SSK study (13, 14) and are between 2 and 14 times the temperature of the MC runs. We also carried out several runs at a temperature of 1.0 and found similar behavior. The temperatures are measured in units of the width of the B_{ij} distribution.

28 contacts and the remaining 128 values randomly to the noncontacts. This ensures that the target structure is the ground state and also leads naturally to a large gap between the ground state and the first excited state. Our scheme results in

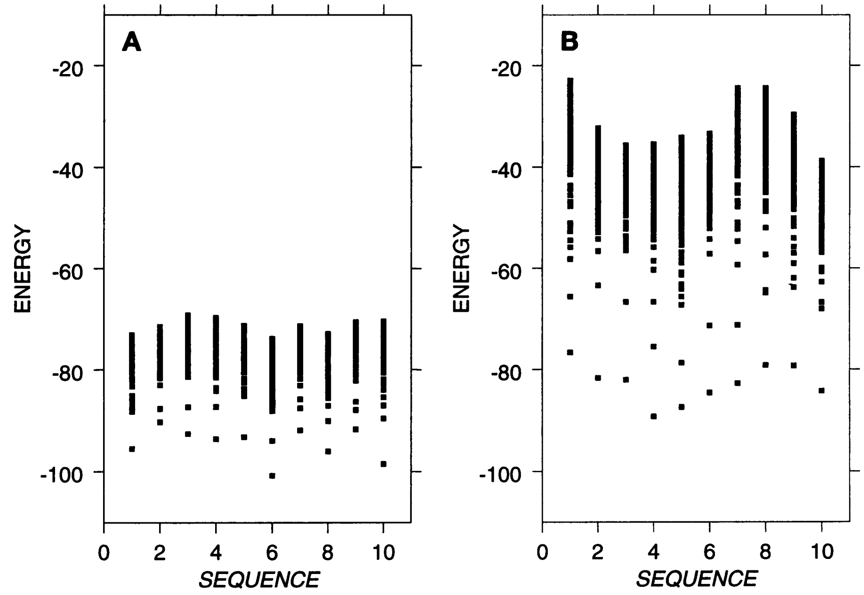

FIG. 1. Energy spectra of the self-avoiding maximally compact conformations for 10 sequences for the purely attractive case (*A*) and for the partially repulsive case (*B*). The 400 lowest energies are shown.

an increase of the correlation between the interaction matrix and the native contact map—the Pearson correlation coefficient (24) is around 0.6 compared with a value around 0.35 for the random sequences of SSK (13, 14) and 1 in an earlier study of Gõ and Abe (25) in which monomers that are in contact in the native state interact attractively while other monomers do not interact at all. Our MC folding simulations [identical to that carried out by SSK (13, 14)] show that the mean foldicity of the designed proteins is 0.95 compared to the 0.13 obtained by SSK (13, 14) for randomly chosen sequences. Thus our choice of maximizing the compatibility of the interacting monomers effectively ensures the stability of the native state over a range of temperatures and allows for kinetic accessibility of this state.

The situation when some of the B_{ij} values are repulsive cannot be investigated within the framework of the original SSK study (13, 14), since the native state for randomly chosen contact energies of either sign (repulsive or attractive) is not necessarily nor likely to be maximally compact. Since an exhaustive search of only the maximally compact conformations is feasible, it is not possible to identify the ground state of the chain. Our simple idea of protein design allows us to assign the strongest attractive interactions to the native contacts thus ensuring that the ground state is maximally compact and known. This scheme is consistent with both maximal compatibility and minimal frustration (8, 18–21). To study this case, we have carried out another set of MC folding simulations in which, for simplicity, the interactions are randomly attractive or repulsive with equal probabilities. Strikingly, now the measured foldicity is 1 and the folding is much faster than in the purely attractive case (Table 1).

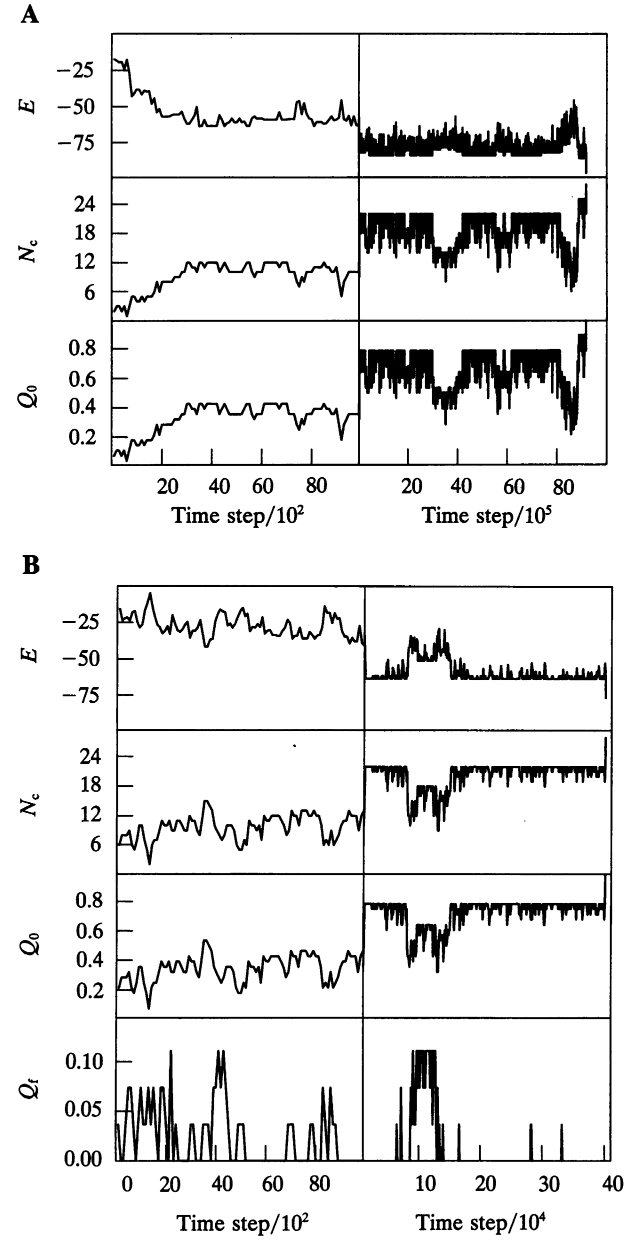

FIG. 2. Typical MC trajectory at T = 1.3 for the purely attractive case (sequence 8) (*A*) and the partially repulsive case (sequence 1) (*B*). The energy, E, is measured in units of T. N_c is the number of contacts, Q_0 denotes the fraction of contacts in common with the native state, and Q_f is the number of repulsive contacts divided by 28. For a maximally compact conformation, N_c = 28.

Fig. 1 demonstrates that the energy gaps between the native and first excited state are indeed large. (Note that while the ground state is exact, the first excited state is obtained just within the set of maximally compact conformations. The true first excited state in the partially repulsive case may not be maximally compact.) The relative stability of the native state among the maximally compact configurations is significantly higher for the partially repulsive case. Typical MC trajectories are shown for both the purely attractive and partially repulsive cases (Fig. 2). In both cases, the dynamics are similar and entail the collapse of the random coil state to a more compact form (due to the attractive interactions). The second stage involves an evolution in the manifold of the semicompact globular state until a rapid folding into the native state occurs. We observed

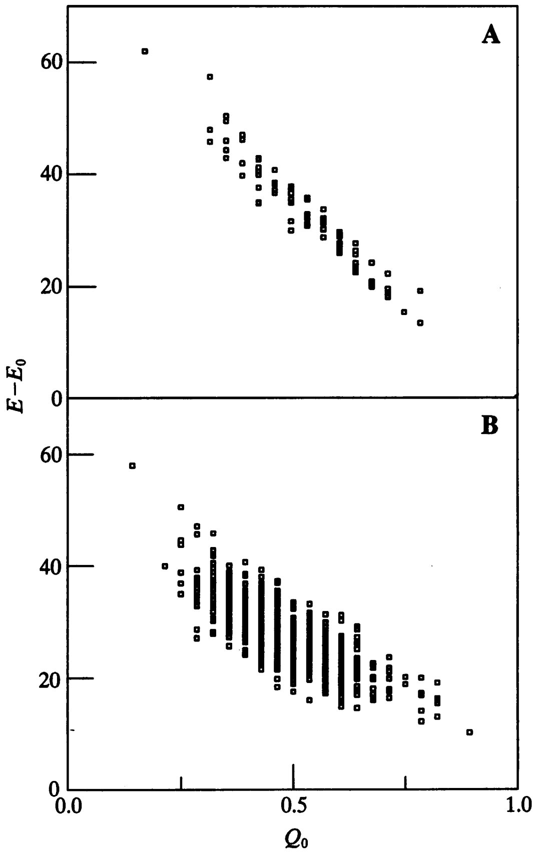

FIG. 3. Plot of ($E - E_0$) vs. Q_0 for the trajectories of Fig. 2. (*A*) The purely attractive case. (*B*) The partially repulsive case. The instantaneous values are plotted every 1000 time steps in *A* and every 44 time steps in *B* so that there are 9162 points in each case (there is a significant overlap of the points). The scatter plots demonstrate the dramatic reduction in the phase space explored on allowing for repulsive interactions.

significant variations in the duration of the second stage (8, 28). The incorporation of repulsion hastens the process of folding by about an order of magnitude (Table 1) due to a significant reduction in the phase space that the chain needs to explore (Fig. 3)—an energetic funnel to the native state has been created. This speeding up is consistent with the increase in the folding transition temperature (8), which is a measure of the thermodynamic stability of the native state, for the partially repulsive case. The number of repulsive contacts remains small (typically ≤3) during most of the MC run (Fig. 2).

Our results may have implications for real proteins. They suggest that sequences of proteins may have evolved in such a way as to minimize the frustration and maximize the compatibility of its elements and, thus, have a large propensity to fold. This may also apply to the structure of RNA and other self-assembly processes. It is interesting to note that a sequence will fold to its unique native state when the most favorable interactions are alloted to the native contacts independent of the precise nature of the assignment (in our model, there are 28! ways of doing this). This suggests that nonhomologous sequences can have the same structure (29–32). Our results raise the possibility that changes in activity and stability of a protein take place through convergent evolution whereas the acquisition of a specific conformation by nonhomologous proteins may have a divergent evolutionary origin.

Our results have implications for protein design and engineering. Extensive experimental studies have shown that it is possible to successfully design proteins that fold into approximately correct structures (33). While protein design is generally carried out at the level of monomers to obtain the desired conformation of a fragment of protein, a few studies have focused on tuning the interactions (34) or generating alternative interactions (35, 36) at specific parts of proteins. The present study is an example of the latter approach and involves the engineering of interactions to design sequences. Further, to achieve the goal of designing unique native structures, the idea of negative design to block certain structures of protein segments is rapidly gaining importance (37). For example, the protein helix has been designed to block unwanted alternatives of reverse topology of helix bundles whereas betabillin excludes edge to edge aggregation of β-sheets (37). Our studies of the reduction of phase space during the folding process of the partially repulsive heteropolymer chains is an example of the general concept of negative design.

In summary, we have worked out the condition for maximum foldicity of polypeptide chains based on a principle of maximum compatibility. We have found that the process of protein folding is speeded up significantly by the presence of repulsive groups, which suggests a solution to the Levinthal paradox. While our studies have been carried out in a well-characterized lattice model, the governing principle ought to be valid more generally and indeed in real proteins. A crucial aspect of our work is that the design of the protein as well as tests of the foldicity have been carried out with the same interaction potentials. [Our scheme is a simpler version than that used by Shakhnovich (38) in a different context.] Our studies have implications in devising algorithms for designing proteins and for prediction of their structure from the sequence information.

We are indebted to Peter Wolynes and an anonymous referee for very helpful suggestions. This work was supported by Department of Bio-Technology (India), U.S. Agency for International Development, North Atlantic Treaty Organization, National Science Foundation, National Aeronautics and Space Administration, the Petroleum Research Fund administered by the American Chemical Society, Office of Naval Research, Komitet Badan Naukowych (Poland), Istituto Nazionale di Fisica Nucleare (Italy), the Fulbright Foundation, Engineering and Physical Sciences Council (U.K.), and grants of computer time by the Science and Engineering Research Council at the Indian Institute of Science and the Center for Academic Computing at The Pennsylvania State University.

1. Creighton, T. E., ed. (1992) *Protein Folding* (Freeman, New York).
2. Levinthal, C. (1969) in *Mossbauer Spectroscopy in Biological Systems*, eds. Debrunner, P., Tsibris, J. C. M. & Munch, E. (Univ. Illinois Press, Urbana), pp. 22–24.
3. Kirkpatrick, S., Gelatt, C. D. & Vecchi, M. P. (1983) *Science* **220,** 671–680.
4. Huberman, B. A. & Hogg, T. (1984) *Phys. Rev. Lett.* **52,** 1048–1051.
5. Unger, R. & Moult, J. (1993) *Bull. Math. Biol.* **55,** 1183–1198.
6. Scheraga, H. A. (1994) *J. Protein Chem.* **13,** 468–469.
7. Rabow, A. A. & Scheraga, H. A. (1993) *J. Mol. Biol.* **232,** 1157–1168.
8. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1995) *Proteins Struct. Funct. Genet.* **21,** 167–195.
9. Gō, N. (1983) *Annu. Rev. Biophys. Bioeng.* **12,** 183–210.
10. Skolnick, J. & Kolinski, A. (1989) *Annu. Rev. Phys. Chem.* **40,** 207–235.
11. Chan, H. S. & Dill, K. A. (1991) *Annu. Rev. Biophys. Biophys. Chem.* **20,** 447–490.
12. Dill, K. A., Bromberg, S., Yue, K., Fieberg, K. M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995) *Protein Sci.* **4,** 561–602.
13. Sali, A., Shakhnovich, E. & Karplus, M. (1994) *Nature (London)* **369,** 248–251.
14. Sali, A., Shakhnovich, E. & Karplus, M. (1994) *J. Mol. Biol.* **235,** 1614–1636.
15. Miyazawa, S. & Jernigan, R. L. (1985) *Macromolecules* **18,** 534–552.
16. Gribskov, M. & Devereux, J., eds. (1991) *Sequence Analysis Primer* (Stockton, New York), Appendix III, p. 229.
17. Iori, G., Marinari, E. & Parisi, G. (1991) *J. Phys. A.* **24,** 5349–5362.
18. Bryngelson, J. D. & Wolynes, P. G. (1989) *J. Phys. Chem.* **93,** 6902–6915.
19. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7524–7528.
20. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 4918–4922.
21. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 9029–9033.
22. Cieplak, M., Maritan, A. & Banavar, J. R. (1994) *Phys. Rev. Lett.* **72,** 2320–2323.
23. Newman, C. M. & Stein, D. L. (1994) *Phys. Rev. Lett.* **72,** 2286–2289.
24. Walpole, R. E. (1974) *Introduction to Statistics* (Macmillan, New York).
25. Gō, N. & Abe, H. (1981) *Biopolymers* **20,** 1013–1031.
26. Chan, H. S. (1995) *Nature (London)* **373,** 664–665.
27. Karplus, M., Sali, A. & Shakhnovich, E. (1995) *Nature (London)* **373,** 665.
28. Socci, N. D. & Onuchic, J. N. (1994) *J. Chem. Phys.* **101,** 1519–1528.
29. Matthews, B. W. (1987) *Biochemistry* **26,** 6885–6888.
30. Flores, T. P., Orengo, C. A., Moss, D. S. & Thornton, J. M. (1993) *Protein Sci.* **2,** 1811–1826.
31. Chothia, C. & Finkelstein, A. V. (1990) *Rev. Biochem.* **59,** 1007–1039.
32. Lim, W. A. & Sauer, R. T. (1989) *Nature (London)* **339,** 31–36.
33. Rees, A. R., Sternberg, M. I. & Wetzel, R., eds. (1992) *Protein Engineering, A Practical Approach* (Oxford Univ. Press, Oxford).
34. Gokhale, R. S., Agarwalla, S., Francis, V. S., Sant, D. V. & Balaram, P. (1994) *J. Mol. Biol.* **235,** 89–94.
35. Ponder, J. W. & Richards, F. M. (1987) *J. Mol. Biol.* **193,** 775–791.
36. Reidhaar-Olson, J. F. & Sauer, R. T. (1988) *Science* **241,** 53–57.
37. Richardson, J. S., Richardson, D. C., Tweedy, N. B., Gernet, K. M., Quinn, T. P., Hecht, M. H., Erickson, B. W., Yan, Y., McClain, R. D., Donlan, M. E. & Surles, M. C. (1992) *Biophys. J.* **63,** 1186–2109.
38. Shakhnovich, E. I. (1994) *Phys. Rev. Lett.* **72,** 3907–3910.