Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 6369–6372, July 1993
Biophysics

Kinetics and thermodynamics of folding in model proteins

CARLOS J. CAMACHO AND D. THIRUMALAI

Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742

Communicated by Robert Zwanzig, March 9, 1993

ABSTRACT **Monte Carlo simulations on a class of lattice models are used to probe the thermodynamics and kinetics of protein folding. We find two transition temperatures: one at T_θ, when chains collapse from a coil to a compact phase, and the other at T_f ($< T_\theta$), when chains adopt a conformation corresponding to their native state. The kinetics are probed by several correlation functions and are interpreted in terms of the underlying energy landscape. The transition from the coil to the native state occurs in three distinct stages. The initial stage corresponds to a random collapse of the protein chain. At intermediate times τ_c, during which much of the native structure is acquired, there are multiple pathways. For longer times τ_r ($\gg \tau_c$) the decay is exponential, suggestive of a late transition state. The folding time scale ($\approx \tau_r$) varies greatly depending on the model. Implications of our results for *in vitro* folding of proteins are discussed.**

It is known that biologically active proteins fold into a native state characterized by a fairly well-defined three-dimensional structure (1). However, the mechanism—the kinetic aspects and pathways—of the folding process remains as one of the most important unsolved problems in structural biology. It is not clear whether the native state corresponds to a global free-energy minimum structure (2) or whether kinetic considerations trap the protein into a metastable minimum (3, 4). The experimental studies of the kinetics of folding have been difficult to carry out; thus the determination of the folding mechanism has remained elusive (5, 6). However, the rebinding kinetics of small ligands to heme proteins following flash photolysis (7, 8) suggest that kinetic events in the folding process should be dictated by the complex energy landscape in proteins (9, 10).

Several authors (11–16) have used random heteropolymer models to gain useful insights into the thermodynamics of protein folding. In this paper, we adopt a different and more natural approach (3, 4, 17–21) by considering three (nonrandom) lattice models of proteins. Our main objective is to understand the dynamics of the folding process in terms of the underlying energy landscape. The most novel aspect of our study is that important features of the energy landscape—namely, barrier heights and the connectivity between low-energy states of the system—have been characterized. Thus, relaxation processes can be interpreted directly in terms of the computed topography of the energy landscape and folding pathways.

Our understanding of macromolecules has been greatly advanced from the study of idealized models. In keeping with such an approach, we have considered a class of simple (d = 2)-dimensional models of proteins. These consist of self-avoiding walks (SAWs) of N sites on the square lattice. The sites in a SAW can only be of two types, hydrophobic (H) or hydrophilic (P). A sequence is specified by the nature of each lattice site (H or P). In our models, interactions involving P sites are taken to be zero. These models embody the dominant interactions in proteins: the flexible connectivity of the residues represented here as sites; self-avoidance arising from steric repulsion between distinct residues; and attractive interactions (potential of mean force) between hydrophobic beads, representing the major driving force of the folding process. Dill and Chan (17, 18) have argued that many of the phenomena observed in proteins can be adequately understood in terms of model A (see below). For all the models, a given structure and its mirror image have identical energies. The energy of a configuration for each of the models is as follows.

A. Heteropolymer Model. Given the coordinates of the SAW sites (r_i, $i = 1, 2, \ldots, N$), the energy of a configuration is determined by the number of topological contacts involving only H sites. Such contacts satisfy the criteria $|i - j| \geq 3$ and $r_{ij} \equiv |\mathbf{r}_i - \mathbf{r}_j| = a$, where a is the lattice spacing. Each hydrophobic contact is assumed to have an attractive energy $-\varepsilon$, with $\varepsilon \geq 0$. The energy of a configuration for this model is given by $E_A = -\varepsilon\Sigma_{i>j}\delta_{r_{ij},a}$ (17, 18), where the sum is over only HH pairs.

B. Bond Saturation Model. The maximum number of favorable nonbonded contacts in the square lattice is three. In model B the decrease in energy when there are multiple nonbonded H contacts is assumed to be nonadditive. Thus, we have taken into account possible packing hindrance. The energy of a chain explicitly depends on the number of HH contacts at each site and is taken to be

$$E_B = -\varepsilon \sum_j g\left(\sum_{i \neq j} \delta_{r_{ij},a} \right), \qquad [1]$$

where $g(0) = 0$, $g(1) = 3/5$, $g(2) = 9/10$ and $g(3) = 6/5$, and the sum is only over HH pairs. Thus, many pairwise interactions are preferred to three- and four-particle interactions.

C. Affine-Sites Model. In proteins the interaction between residues can be quite varied depending on the location as well as the nature of residues in the primary sequence. This specificity is mimicked by dividing the H sites into noninteracting subsets. As a further restriction we have chosen the subsets among the sets of topological contacts found in the ground state of models A and B (see, e.g., Fig. 1) (19, 20). The allowed interactions between H sites belonging to the same subset are the same as in model A.

For each model, we have generated random sequences of H and P sites for $N = 15$ and the number of H sites $N_H = 8$. We have found—using *exact* series enumerations of $d = 2$ SAWs—that only 3.2% of all possible sequences lead to a unique ground state for model A, whereas the corresponding percentage for model B is 7.6%. Since we are primarily interested in the kinetics of approach to the native state (believed to be unique in proteins), we have considered only those sequences of H and P residues that lead to a nondegenerate ground state (except for the mirror image). Numerical results presented here are restricted to those obtained for the sequence described in Fig. 1. Qualitatively similar results

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: SAW, self-avoiding walks; MC, Monte Carlo; H, hydrophobic; P, hydrophilic.

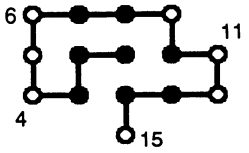

FIG. 1. Ground-state configuration for the sequence shown. $N = 15$; ●, H; ○, P. The total number of distinct structures is 296,806. For model C, the subsets of interacting H sites are the following {1, 8, 10, 14}, {2, 7}, {3, 14}, and {10, 13}. Notice that these interactions involve only nearest neighbor contacts present in the ground state.

have been found for other sequences as well. The simulation results have been obtained by using single-site lattice Monte Carlo (MC) dynamics (22) and a standard Metropolis algorithm. The energy scale, ε, the Boltzmann constant, k_B, and the lattice spacing, a, have been set equal to unity.

The energy levels and the associated degeneracies have been explicitly enumerated, from which the thermodynamic properties of the folding process can be *exactly* calculated. On the other hand, the dynamical properties are determined not only by the spectrum but also by the connectivity of the energy states. However, given the multidimensional nature of this problem, an accurate representation of the energy landscape connecting different states is in general difficult to construct. Nevertheless, due to the discreteness of our models, we have been able to characterize the connectivity of at least some of the relevant low-energy states. Fig. 2 displays energy landscapes indicating the minimal energy pathways between adjacent states for models A–C in the neighborhood of the ground state (S); i.e., every connecting pathway between two local minima has to go over an energy barrier of at least the height indicated on the figure. Each well contains a number of slightly different overall structures connected to the local minimum by at least one monotonically decreasing energy pathway. We have computed the total number of distinct structures belonging to a given well (at or below the reference energy represented by the dashed lines in Fig. 2) by an extensive MC search (see Fig. 2 for some examples).

The folding kinetics has been probed by studying different relaxation functions as a function of time—i.e., MC steps. These functions include the radius of gyration, $\langle R_G^2(t)\rangle = \frac{1}{2N^2}\langle \Sigma_{i,j} r_{ij}^2(t)\rangle$; the energy, $\langle E(t)\rangle$; and the probability that a sequence is in its ground state, $\langle P(t)\rangle = \langle \delta[\Sigma_{i,j}\{r_{ij}(t) - r_{ij}^0\}]\rangle$, where the superscript 0 refers to the ground state. As a further microscopic probe of the folding process we have also computed an "overlap" function χ (≤ 1) that measures the degree of disorder of the protein structure with respect to the ground state,

$$\langle \chi(t)\rangle = 1 - \frac{1}{N^2 - 3N + 2}\left\langle \sum_{i\neq j,j\pm 1} \delta(r_{ij}(t) - r_{ij}^0)\right\rangle. \quad [2]$$

The overlap function χ, which is zero only when the protein is in its ground-state configuration, can be thought of as a measure of the amount of non-native-like structure present in the protein. To address the nature of the folding transition we have also computed fluctuation quantities such as the specific heat and the fluctuations in the overlap function, $\Delta\chi = \langle\chi^2\rangle - \langle\chi\rangle^2$. Note that $\Delta\chi = 0$ implies that the system is in a pure state (namely the ground state for our models) and in this sense $\Delta\chi$ is analogous to susceptibility in magnetic systems.

The folding dynamics has been inferred by studying the relaxation to equilibrium of a set of independent infinite temperature conformations after a temperature quench. The aforementioned relaxation functions have been calculated by averaging over many independent runs ranging from 900 to 20,000, depending on temperature. Our numerical experiments were restricted to a finite amount of computational time, $\tau_\infty \approx 4 \times 10^{10}$ MC steps, thus limiting the value of N to 22.

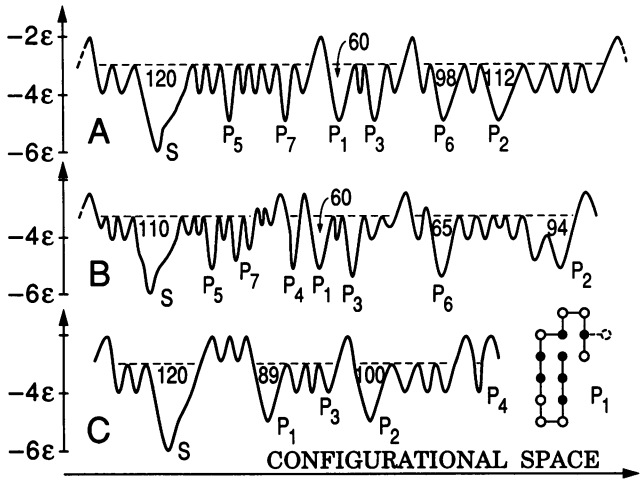

FIG. 2. Energy landscape for models A–C. Every configuration present at or below the dashed line has been enumerated (a total of 783, 542, and 479, respectively). The landscape for model C includes all the states with $E_C < -4\varepsilon$. The grouping of states is unique but at the energy level represented by the dashed lines. The ordering of the different wells on the horizontal axis is roughly determined by how many MC steps are needed to reach the native well.

Before describing the dynamical behavior, let us discuss the relevant thermodynamic transitions that take place in these models as the temperature is lowered. High-temperature conformations behave very much like SAWs or polymers. At a lower temperature T_θ, the chain undergoes a continuous collapse transition into more compact configurations (e.g., ref. 23 and references therein). The temperature T_θ can be estimated by computing the energy fluctuations in the system, where the resulting plot of the specific heat as a function of T shows a peak. From the specific-heat plots we have estimated $T_\theta \approx 0.39$, 0.34, and 0.41 for models A, B, and C, respectively.

It is natural to assume that the transition to the native state takes place from the set of compact structures found at $T < T_\theta$. To ascertain whether an additional amount of native structure is acquired with a further decrease in T, we have calculated $\langle\chi\rangle$ as well as $\Delta\chi$ as a function of temperature. As T is lowered, a rather abrupt decrease of $\langle\chi\rangle$ is observed (unpublished work), suggesting the possibility of an additional transition. Further evidence for this transition emerges when the temperature dependence of $\Delta\chi$ is analyzed. The peaks observed in Fig. 3 occur at the same temperature at which the abrupt change in $\langle\chi\rangle$ takes place. We have associated these peaks with a folding transition temperature T_f,

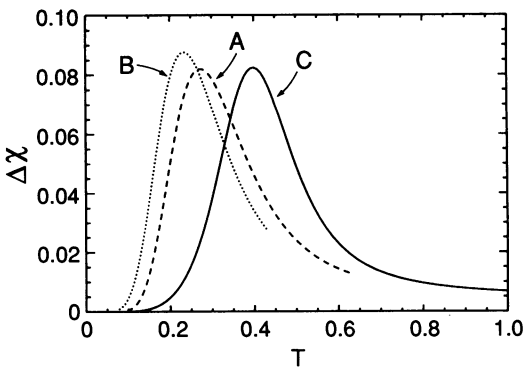

FIG. 3. $\Delta\chi$ as a function of temperature for models A–C.

where $T_f \approx 0.27$, 0.23, and 0.40 for models A, B, and C, respectively. It is reasonable to suggest that the conformations in proteins in the region $T_f \leq T \leq T_\theta$ may correspond to a "molten globule" containing the bulk of the backbone native structure. It is noteworthy that a mean-field replica analysis of a model similar to model C also shows a molten-globule phase (11–13), suggesting that this transition may be generic for proteins.

The origin of the transition at $T = T_f$ can be understood in terms of the energy landscape shown in Fig. 2. At low temperatures, the free energy can be envisaged as having essentially two wells: one well containing the native state, and the other a collection of low-energy states. The transition temperature T_f indicates the possible coexistence between these two distinct states. By examining the time dependence of $\langle \chi(t) \rangle$ at $T \approx T_f$, coexistence—defined as transitions between native structure ($\langle \chi \rangle \approx 0$) and a distinctly different type of structure (mostly the P levels with $\langle \chi \rangle \approx 0.4$)—is observed (unpublished work). Thus, it can be concluded that the anomalous fluctuations present in the protein structure at $T \approx T_f$ are indicative of finite-size, first-order transitions.

All the models exhibit similar relaxational behavior, although the temperature scales are different (see below). Typical plots of $\langle \chi(t) \rangle$ and $\langle R_G^2(t) \rangle$ for model C and $T = 0.286$ ($< T_f^C$) are shown in Fig. 4 *a* and *b*, respectively. These figures also show the striking contrast between the relaxation functions averaged over several initial conditions and that obtained in a single realization (dotted lines). The curvature in $\langle \chi(t) \rangle$ and $\langle R_G^2(t) \rangle$ is a clear signature of the multistage character of the folding process. [It has been argued (3, 4, 21) that folding proceeds by a two-stage process.] Although one can interpret the behavior found in $\langle \chi(t) \rangle$ and $\langle R_G^2(t) \rangle$ in terms of a distribution of time scales, here we will focus on two of them, denoted by τ_c and τ_r.

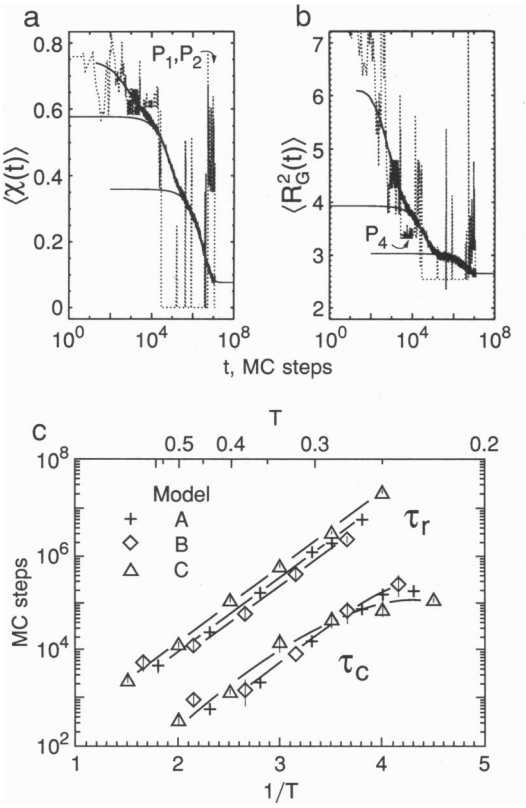

FIG. 4. (*a* and *b*) $\langle \chi(t) \rangle$ and $\langle R_G^2(t) \rangle$ as a function of time for model C and $T = 0.286$. The ensemble average has been done for 2000 independent runs, an example of which is represented by the dotted line. Some intermediate states are also indicated. The smooth solid lines shown in both *a* and *b* represent the best biexponential fit for the two final stages of folding. (*c*) Temperature dependence of τ_r and τ_c (the latter is shifted to the right by 0.5). Lines are a guide to the eye.

For $t \lesssim \tau_c \approx 6 \times 10^4$ MC steps, proteins reach a high degree of compactness [as measured by $\langle R_G^2(t) \rangle$]. As indicated by the large decrease in $\langle \chi(t) \rangle$, much of the native structure is formed in this intermediate time scale (3, 4); thus, we conclude that this stage of the folding process is highly cooperative. A noncooperative process, related to the initial random collapse of the protein chain, has also been observed for $t \ll \tau_c$ (see Fig. 4 *a* and *b* for $t \lesssim 10^3$ MC steps). For $\tau_c < t \lesssim \tau_r \approx 3 \times 10^6$ MC steps, the folding kinetics show a much slower decrease of $\langle \chi(t) \rangle$ to its equilibrium value $\langle \chi \rangle = 0.07706$.

We have found that in the final stage of the folding process $\langle \chi(t) \rangle$ and $\langle R_G^2(t) \rangle$ decay exponentially and are well characterized by a single decay constant, τ_r. For shorter times the decay process is more complex, indicating that the system may in fact be sampling several distinct structures that are separated by varying barrier heights. In this time regime the decay of $\langle \chi(t) \rangle$ could be nonexponential. Indeed, a stretched exponential fit {i.e., $\langle \chi(t) \rangle \approx \exp[-(t/\tau)^\beta]$} is consistent with our results for models A–C. However, our data are not precise enough to unambiguously extract the exponent β. Hence, we have assumed that the intermediate time decay is also exponential and have extracted a time constant τ_c for this regime. Biexponential fits to $\langle \chi(t) \rangle$, $\langle P(t) \rangle$, $\langle R_G^2(t) \rangle$, and $\langle E(t) \rangle$ have been done; to accomplish this we have used the exact thermodynamic values for $\langle \chi \rangle$ and $\langle R_G^2 \rangle$. The results are summarized in Fig. 4*c*, where the temperature dependence of τ_r and τ_c have been plotted. The solid lines in Fig. 4 *a* and *b* represent the biexponential fit.

The relationship between kinetics and folding pathways can be analyzed in terms of the relaxation functions and the underlying energy landscape shown in Fig. 2. We have studied several individual runs and have observed that there are multiple pathways to the ground state in the first two stages of the folding process ($t \lesssim \tau_c$) (3, 4, 24). In Fig. 4 *a* and *b*, for instance, the initial transition to the S state takes place from the P_4 conformation; other realizations (not shown) include transitions from P_5, P_6, or P_7 to S (see Fig. 2). However, pathways which early on in the folding process include P_1 and/or P_2 relax to the ground state on a time scale of the order of τ_r. In all cases we find that P_1 and/or P_2 are always involved in the late stage of folding and represent the major bottleneck for the transition to the native state. Fig. 4 *a* and *b* show that transitions to nearby metastable states [e.g., $S \rightarrow (P_1, P_2) \rightarrow S$] occur even after proteins have reached the ground state; notice that the lifetime of this transition is of the order of τ_r. These transitions involve a rearrangement of the whole protein structure and nonbonded contacts. Indeed, as indicated by the large fluctuations observed in $R_G^2(t)$, there is considerable unraveling of the protein even in the late stages of folding.

A striking result of our study is that the degree of compactness and the amount of native structure present do not have a one-to-one correspondence. Indeed, for $T \gtrsim T_\theta$, models A and B are significantly more compact than model C. Furthermore, for $T \lesssim T_\theta$ and $t \lesssim \tau_r$, the rate of decrease of $\langle R_G^2(t) \rangle$ is also much larger for models A and B than for model C, and a similar degree of compactness is reached only in the final stage of folding. However, for models A and B, the rate of decrease of $\langle \chi(t) \rangle$ is smaller than for model C for all t. Moreover, the folding kinetics show that during the early folding process ($t \leq \tau_c$), the amount of native structure is roughly 4 times larger in model C than in models A and B. The folding time at $T = T_f$ estimated from Fig. 4*c* is 5×10^6, 3×10^7, and 10^5 MC steps for models A, B, and C, respectively, which is consistent with the above observations. Therefore

searching for the native state within the space of compact configurations does not imply that the folding time scales are equivalent. In fact, they are determined by the dynamics of $\langle\chi(t)\rangle$. There appears to be a relationship between the folding time τ_r and $\sigma = [1 - (T_f/T_\theta)]$: the smaller the value of σ, the smaller the value of τ_r. This relationship may prove quite helpful in designing sequences that fold in finite times.

In order to use our findings to predict the experimental time scales for τ_c and τ_r, it is necessary to understand the N dependence of these relaxation times. It is difficult to unambiguously ascertain the N dependence in systems with heterogeneous interactions. Nevertheless, one may argue that τ_c (the time scale related to the compactness process) should scale like the relaxation time for homopolymers in a poor solvent—i.e., $\tau_c \approx N^z$, with $z \approx 2$–3 being roughly the dynamical scaling exponent (25). Our limited data are consistent with this behavior. Thus, for a protein with $N = 58$ (pancreatic trypsin inhibitor), we find that τ_c taken from Fig. 4 *a* and *b* extrapolates to $\tau_c \approx 10^{-4}$–10^{-3} s, where an estimate of 10^{-11}–10^{-10} s has been used to estimate one MC step. It is harder to produce theoretical arguments for the scaling of τ_r with N because this would require estimates of entropy of activation. However, we believe that our results for the energy landscapes may be a good starting point from which to advance a solution to this problem.

The dynamical results presented here have implications for the kinetics of *in vitro* folding of proteins. (*i*) Our results suggest that the basic kinetic scheme for approach to the native state—namely, the acquisition of the bulk of the secondary structure via multiple pathways in an intermediate time scale τ_c, followed by an activated crossing over a late transition state in times almost 2 orders of magnitude greater than τ_c—may be a robust feature of all sequences that have a native state. The relevance of τ_c is that most of the process of becoming compact and the formation of secondary structure have taken place within this relatively short time scale. From an evolutionary point of view, it is tempting to speculate that only those sequences for which τ_c is short (10^{-4}–10^{-3} s in proteins) will fold in a biological time scale (seconds). Furthermore, the unmistakenly exponential decay of $\langle\chi(t)\rangle$ at long times for all the models, along with the similar value of the effective barrier height of 3.6ε obtained from the temperature dependence of τ_r in Fig. 4*c* (consistent with 4ε being the highest energy barrier in Fig. 2), suggests that the nature of the transition state is essentially the same in all cases. However, the fast folding time scale for model C implies that the approach to the native state is greatly expedited if the residues are in their ground state (26, 27). (*ii*) A major surprise in our study is that energy landscapes that are far simpler than that expected in proteins can in fact lead to highly complicated dynamics. These findings provide strong support for the use of simple lattice models in yielding helpful insights into the dynamics of real proteins.

The conclusions we have drawn for the folding kinetics of proteins should be viewed with caution, however. Side chains, which are responsible for hydrogen bonding and other tertiary interactions, have not been considered here. Despite this limitation and additional restrictions of lattice models, we have advanced a testable kinetic scheme for protein folding which should be useful in understanding fast folding experiments currently underway in several laboratories.

We are grateful to E. Shakhnovich for helpful correspondence; C.J.C. thanks D. Kandel, C. Doty, A. Boutaud, and H. Rieger for valuable discussions; and D.T. is grateful to Bill Eaton for several illuminating discussions and comments on the manuscript. This work was supported in part by a grant from the National Science Foundation.

1. Stryer, L. (1988) *Biochemistry* (Freeman, San Francisco).
2. Anfinsen, C. B. (1973) *Science* **181,** 223–230.
3. Honeycutt, J. D. & Thirumalai, D. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 3526–3529.
4. Honeycutt, J. D. & Thirumalai, D. (1992) *Biopolymers* **32,** 695–709.
5. Weissman, J. S. & Kim, P. S. (1991) *Science* **253,** 1386–1393.
6. Creighton, T. E. (1992) *Nature (London)* **356,** 194–195.
7. Iben, I. E. T., Braunstein, D., Doster, W., Fraunfelder, H., Hong, M. K., Johnson, J. B., Luck, S., Ormos, P., Schulte, A., Steinbach, P. J., Xie, A. H. & Young, R. D. (1989) *Phys. Rev. Lett.* **62,** 1916–1919.
8. Tian, W. D., Sage, J. T., Srajer, V. & Champion, P. M. (1992) *Phys. Rev. Lett.* **68,** 408–411.
9. Frauenfelder, H., Parak, F. & Young, R. D. (1988) *Annu. Rev. Biophys. Chem.* **17,** 451–479.
10. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. (1991) *Science* **254,** 1598–1603.
11. Shakhnovich, E. & Gutin, A. M. (1989) *Studia Biophys.* **132,** 47–56.
12. Shakhnovich, E. & Gutin, A. M. (1989) *Biophys. Chem.* **34,** 187–199.
13. Shakhnovich, E. & Gutin, A. M. (1989) *J. Phys. A* **22,** 1647–1659.
14. Garel, T. & Orland, H. (1988) *Europhys. Lett.* **6,** 307–310.
15. Garel, T. & Orland, H. (1988) *Europhys. Lett.* **6,** 597–601.
16. Kantor, Y. & Kardar, M. (1991) *Europhys. Lett.* **14,** 421–426.
17. Dill, K. A. (1990) *Biochemistry* **29,** 7133–7155.
18. Chan, H. S. & Dill, K. A. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 6388–6392.
19. Taketomi, H., Ueda, Y. & Gō, N. (1975) *Int. J. Pept. Protein Res.* **7,** 445–459.
20. Ueda, Y., Taketomi, H. & Gō, N. (1978) *Biopolymers* **17,** 1531–1548.
21. Shakhnovich, E., Farztdinov, G., Gutin, A. M. & Karplus, M. (1991) *Phys. Rev. Lett.* **67,** 1665–1668.
22. Verdier, P. H. (1973) *J. Chem. Phys.* **59,** 6119–6127.
23. Kremer, K., Baumgartner, A. & Binder, K. (1981) *J. Phys. A* **15,** 2879.
24. Skolnick, J. & Kolinski, A. (1989) *J. Mol. Biol.* **212,** 787–817.
25. de Gennes, P. G. (1979) *Scaling Concepts in Polymer Physics* (Cornell Univ. Press, New York).
26. Bryngelson, J. D. & Wolynes, P. G. (1989) *J. Phys. Chem.* **93,** 6902–6915.
27. Zwanzig, R., Szabo, A. & Bagchi, B. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 20–22.