UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL ATOMIC ENERGY AGENCY
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

H4.SMR/916 - 30

SEVENTH COLLEGE ON BIOPHYSICS:

Structure and Function of Biopolymers: Experimental and Theoretical Techniques.

4 - 29 March 1996

Protein Folding - 2

Jose N. ONUCHIC
Department of Physics
University of California at San Diego
9500 Gilman Drive
CA 92093 La Jolla
U.S.A.

Reprint Series
17 March 1995, Volume 267, pp. 1619–1620

SCIENCE

Navigating the Folding Routes

Peter G. Wolynes, Jose N. Onuchic, and D. Thirumalai

Navigating the Folding Routes

Peter G. Wolynes, Jose N. Onuchic, D. Thirumalai

To fold, a protein navigates with remarkable ease through a complicated energy landscape as it explores many possible physical configurations. This feat is beginning to be quantitatively understood by means of statistical mechanics and simplified computer models (*1*). Folded proteins are marvels of molecular engineering and it is hard to avoid thinking that all of their complex structural features play a role in their folding through an obligate multistep mechanism. A unique folding pathway, if it exists, could be elucidated with classical chemical experiments. A newer view holds that in the earlier stages a protein possesses a large ensemble of structures. The problem is not to find a single route but to characterize the dynamics of the ensemble through a statistical description of the topography of the free-energy landscape. Folding is easy if the landscape resembles a many-dimensional funnel leading through a myriad of pathways to the native structure. Only a few parameters should be needed to characterize statistically the topography of and routes down the folding funnel. Using experimental data, Onuchic *et al.* have estimated the extent, ruggedness, and slope of the folding funnel (*2*). Similar parameters characterize the energy landscape of simple computer models of proteins. These models of self interacting necklaces of beads, often on lattices, lack most of the details of real proteins, but establishing a quantitative correspondence between the landscapes of computer models and real proteins makes it possible to use simulations to understand folding kinetics.

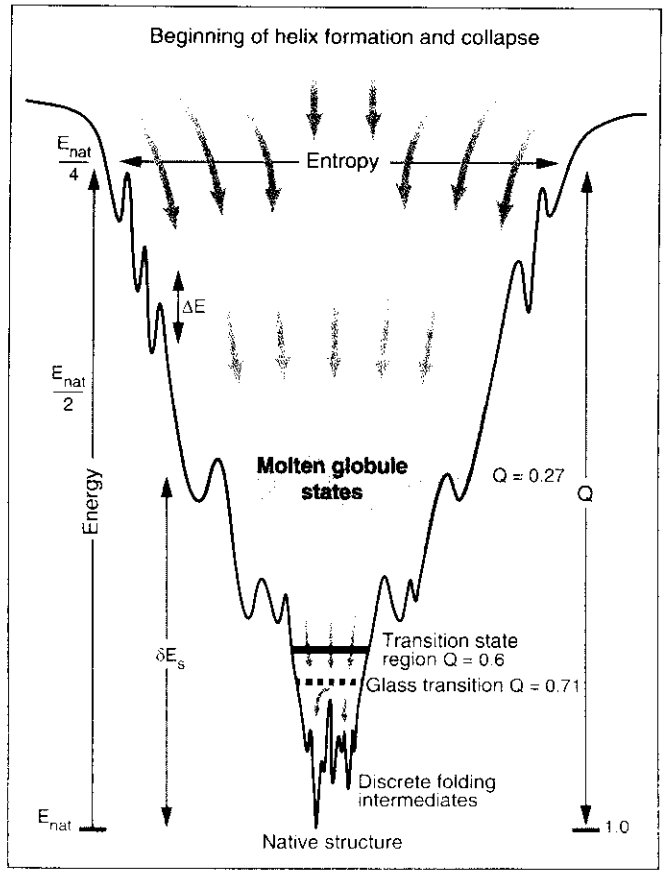


Fig. 1. Schematic of the folding funnel for a fast-folding 60-residue helical protein according to Onuchic *et al.* (*2*). The width of the funnel represents entropy, and depth, the energy. The flow of the molecule through the molten globule, folding bottleneck, or transition state ensemble and a glass transition region where discrete pathways emerge are indicated. The fraction of native contacts correctly made, *Q*, is indicated for each collection of states.

The extent of a protein energy landscape is huge. Before folding, each residue can take on about 10 different conformations; thus, a 60-residue protein can be in any of 10^{60} states. An unguided search, like a drunk playing golf, would take practically forever. A flat energy landscape (or golf course) is very unrealistic, but many years ago Bryngelson *et al.* pointed out that a difficult search also arises on a rugged energy landscape that might describe proteins (*1*). Such a landscape can arise because different segments of a typical heteropolymer come together without any guarantee that the many resulting individual interactions will not mutually conflict and thereby "frustrate" minimizing each other. This landscape will be rugged with many deep valleys corresponding to local minima. Transient trapping in these valleys slows the exploration of routes toward the most stable native structure. This trapping resembles the way a liquid becomes a glass when cooled, remaining fixed in one of many structures, unable to reconfigure to the lowest energy crystal state. At temperatures far above a glass transition, a rough landscape is easily traversed. At low temperatures, where the ground state has a significant chance to be thermally occupied, the search to find the deepest valley out of the many on the glassy energy landscape is incredibly slow. For a protein to be kinetically foldable, there must be a sufficient overall slope of the energy landscape so that the numerous valleys flow in a funnel toward the native structure. With such a slope, native structure elements are significantly more stable than expected by chance. Thus, the global energy minimum (native structure) is still thermodynamically stable above the glass transition temperature, where kinetic barriers for escaping glassy traps (misfolded structures) become too large. That the interactions of a kinetically foldable protein must have fewer conflicts than typically expected is known as the "principle of minimal frustration." Minimizing frustration or the ratio of glass to folding temperatures is equivalent to maximizing the "stability gap" between the native state and disordered collapsed structures measured in units of the ruggedness. The quantitative version of the minimal frustration principle has been used to infer energy functions useful for structure prediction (*3*) and to design proteins in machina (*4*). Many simulations of simple models have confirmed the principles of the energy landscape analysis (*5*), including a recent exhaustive study of ~200 short sequences (*6*).

Onuchic *et al.* look at real protein folding by developing a law of corresponding states (*2*). Studying phase transitions teaches that boiling is very similar for systems as different as water and methane because each system can be mapped onto the same pat of a universal phase diagram. Similarly, Onuchic *et al.* argue that the detailed local structural features like the hydrogen bonds of helices and side chain packing can be taken into account by finding appropriate values of the statistical parameters characterizing the free energy landscape. Helix formation and collapse solve part of the folding search locally so as to renormalize the effective number of degrees of freedom and change the details of single reconfiguration events. Nevertheless, the global features of the landscape of a real protein can be mapped onto those of a bead model.

P. G. Wolynes is in the School of Chemical Sciences, University of Illinois, Urbana, IL 61801, USA. J. N. Onuchic is in the Department of Physics, University of California at San Diego, La Jolla, CA 92093, USA. D. Thirumalai is at the Institute for Physical Sciences and Technology, University of Maryland, College Park, MD 20742, USA.

Fig. 2. (Left) Ribbon diagram for the native structure of a 96-residue helical protein that corresponds with a 46-residue bead model simulated by Thirumalai and Guo (*7*). **(Right)** A discrete native-like intermediate whose rearrangement is the slowest process in their simulation.

A theory of helix formation in collapsed polymers relates the measured amount of helix in the collapsed states to the conformational entropy associated with different chain topologies. The extent of the renormalized configurational energy landscape for a 60–amino acid helical protein is only a bit larger than that for models containing only 27 beads. At the folding temperature, using thermodynamics, the resulting entropy estimate yields the slope of the folding funnel. Finally, measurements on motions in the collapsed denatured proteins quantify the ruggedness and with the conformational entropy leads to values of a glass transition temperature. These estimates suggest real proteins resemble bead models in which only three kinds of residues are used to encode sequence in that both have a folding to the glass transition temperature ratio of 1.6.

A simulation of a three-letter code, 27-mer lattice model gives a picture of the folding mechanism and the folding funnel topography of a 60–amino acid helical protein. A caricature of the multidimensional funnel accurately representing the entropy and structural similarity involved in the search is shown in Fig. 1.

The funnel illustrates that a fast-folding helical protein has a collapsed molten globule band of states with roughly one-quarter of the number of native contacts correctly made. As one proceeds down the funnel, both entropy and energy decrease, but when roughly three-fifths of the native contacts are made, the incomplete compensation of entropy decrease by energy decrease leads to a very modest barrier ($\sim 3k_BT$). Folding is thereby slowed by a thermodynamic bottleneck. The transition state or bottleneck region consists of a large ensemble of structures reflecting the multiple pathways of protein folding. After this bottleneck is crossed, the entropy still decreases until discrete kinetic intermediates appear, most having roughly three-quarters of the correct contacts. These native-like but misfolded structures are sensitive to sequence mutation. If the thermodynamic bottleneck is not too narrow and the landscape still rough, the search through intermediates becomes rate-limiting.

A recent off-lattice study of Thirumalai and Guo (*7*) illustrates this well into the folding of a model with 46 beads. Using the connection between simplified bead models and helical proteins, their simulation is roughly like that of a 96-residue four-helix bundle protein whose native structure is shown in Fig. 2. The sequence design of their model gives a folding funnel with a smaller slope than the fast folders just discussed. Now trapping in a rather native-like intermediate (Fig. 2) becomes a distinct, slow side reaction, while a fraction of molecules follow a more direct nucleation-like mechanism with a small barrier consistent with a finite size scaling estimate.

The correspondence theory puts simulations into direct contact with recent experiments. Several studies have demonstrated that the discrete intermediates in the refolding of cytochrome c can entirely disappear upon changing an individual residue's chemistry (*8*). In keeping with the energy landscape theory, these results show that the discrete intermediates often found are an epiphenomenon, their distinctness unimportant when trying to understand how folding is guided. Intermediates are relics of the landscape ruggedness whose features may fossilize some of the relevant guiding forces. Fersht's recent study of linear free energy relations for folding chymotrypsin inhibitor (*9*) suggests that different residues are between 30 and 70% folded in the ensemble of structures representing the transition state, reminiscent of the bottleneck in the three-letter-code funnel.

The small size of the actual thermodynamic barrier to folding is perhaps a bit surprising. Many studies on long time scales show very large activation barriers, but the landscape theory suggests that these arise from transient trapping. In vivo proteins are stable by several k_BT, so that folding may be largely a downhill run toward the final near native kinetic traps followed by a short search through them, as in many simulations. The near perfect compensation of entropy by enthalpy in the funnel suggests that proteins behave like fluids near a critical point. For theoreticians, a next step is to see how scaling and renormalization group ideas might be used to understand kinetics, especially for larger proteins. A more complete experimental characterization of the dynamics of partially folded proteins throughout their phase diagram, including very low temperatures and high pressures (*10*), is also needed to precisely quantify the glass transition. For experimentalists, the present perspective also shows that the guiding forces act in much less than a few milliseconds. A new generation of experiments using lasers to rapidly initiate folding (*11*) promises dramatic advances in direct measurement of the protein's energy landscape topography.

References and Notes

1. J. D. Bryngelson *et al.*, *Proteins: Struct. Funct. Genet.*, in press, and references therein.
2. J. N. Onuchic *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
3. R. A. Goldstein *et al.*, *ibid.* **89**, 4918 (1992).
4. E. I. Shakhnovich, *Phys. Rev. Lett.* **72**, 3907 (1994).
5. J. D. Honeycutt and D. Thirumalai, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 3526 (1990); C. J. Camacho and D. Thirumalai, *ibid.* **90**, 6369 (1993); R. Miller *et al.*, *J. Chem. Phys.* **96**, 768 (1992).
6. A. Sali *et al.*, *J. Mol. Biol.* **235**, 1614 (1994).
7. D. Thirumalai and Z. Guo, *Biopolymers* **35**, 137 (1995).
8. G. Elove *et al.*, *Biochemistry* **33**, 6925 (1994); T. R. Sosnick *et al.*, *Nature Struct. Biol.* **1**, 149 (1994).
9. D. E. Otzen *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 10422 (1994).
10. J. Jonas and A. Jonas, *Annu. Rev. Biophys. Biomol. Struct.* **23**, 287 (1994).
11. C. M. Jones *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11860 (1993).

12. We thank N. Socci, Z. Luthey-Schulten, and R. Cole for the preparation of the figures and W. Eaton for reading the manuscript. The work of the authors was supported by the National Science Foundation and the National Institutes of Health.

Kinetic and thermodynamic analysis of proteinlike heteropolymers: Monte Carlo histogram technique

Nicholas D. Socci and José Nelson Onuchic
Department of Physics, University of California at San Diego, La Jolla, California 92093-0319

(Received 14 March 1995; accepted 13 June 1995)

Using Monte Carlo dynamics and the Monte Carlo histogram method, the simple three-dimensional 27 monomer lattice copolymer is examined in depth. The thermodynamic properties of various sequences are examined contrasting the behavior of good and poor folding sequences. The good (fast folding) sequences have sharp well-defined thermodynamic transitions while the slow folding sequences have broad ones. We find two independent transitions: a collapse transition to compact states and a folding transition from compact states to the native state. The collapse transition is second-order-like, while folding is first-order-like. The system is also studied as a function of the energy parameters. In particular, as the average energetic drive toward compactness is reduced, the two transitions approach each other. At zero average drive, collapse and folding occur almost simultaneously; i.e., the chain collapses directly into the native state. At a specific value of this energy drive the folding temperature falls below the glass point, indicating that the chain is now trapped in local minimum. By varying one parameter in this simple model, we obtain a diverse array of behaviors which may be useful in understanding the different folding properties of various proteins. © *1995 American Institute of Physics.*

I. INTRODUCTION

Simple models are one powerful theoretical tool for the study of complex systems. The idea is to remove all but the essentials from the original system which will ideally allow for a more complete analysis. There is often a trade-off between the complexity of the model (or how faithfully it represents the system of interest) and the thoroughness of the analysis. In the case of protein folding, research has spanned the entire spectrum from all-atom molecular dynamics with solvent[1–4] to complete enumeration of simple lattice polymer systems,[5,6] with many works in between these two extremes. Naturally the more realistic simulations do not yield results as thorough as the simpler ones. In the all-atoms simulations a large amount of supercomputer time is required for runs of hundreds of picoseconds of a single protein molecule (plus solvent). In contrast, high-end workstations can be used to simulate lattice polymers. Many different sequences can be simulated over a range of temperatures for time scales comparable to the folding time. We do not want to imply that one set of techniques is superior to the other, but rather in studying a system as complex as proteins many different approaches are necessary. In fact the two limits complement each other. Simple systems permit detailed analysis while the more complex systems allow for contact with real proteins. Connecting these two limits would allow for a more thorough analysis of real protein systems. Such an analysis has been recently carried out.[7,8]

In a previous work[9] we examined the kinetics of a simple three-dimensional lattice polymer system. This system is too large for exact enumeration studies but is still small enough for detailed analysis. Many studies of lattice models seem to focus either on thermodynamics or on kinetics, considering each in isolation. However, as previously shown[9] and shown here, a combined approach that considers both the kinetics and thermodynamics of the same system is important in understanding the model in detail. We have determined that there is an important relation between kinetics (the "glass transition") and thermodynamics (the folding temperature) in determining whether a sequence will fold or not, an idea that was advanced by Bryngelson and Wolynes,[10,11] and later explored by Leopold and Onuchic.[12] In this work we continue the study of this system, examining both thermodynamics and kinetics in greater detail.

Some of the earliest work on the thermodynamics of proteinlike lattice polymers has been performed by Chan, Lau and Dill.[13,14] They examined short chains in two dimensions for which it is possible to enumerate all conformations. This allowed them to calculate any thermodynamic quantity by simply summing over states. By measuring a variety of parameters, such as the average compactness, number of contacts and hydrophobic core, they found a distinct difference between folding and nonfolding sequences. Although exact enumeration studies are extremely powerful, they are limited to small chains, usually in two dimensions. In three dimensions many have studied the 27 monomer system on a simple cubic lattice which has a maximally compact shape of a $3\times3\times3$ cube. It is not possible to enumerate all conformations, but it is still possible to enumerate all *compact* conformations (what is often referred to as the *cube spectrum*) and then calculate approximate thermodynamics using just the cube states.[15,16] However, as we show in this work, care must be taken in using this technique since the cube states are not an accurate approximation to the full density of states; in particular there are many low energy (i.e., thermodynamically relevant) states that are not cubes.

For longer chains, one can use the standard Monte Carlo technique[17] for calculating thermodynamic properties. Skolnick, Sikorski and Kolinski[18–20] use a dynamic Monte Carlo method to study folding of realistic proteinlike structures on a diamond lattice. The word dynamic is used to indicate that the move set has been selected in such a way

 0021-9606/95/103(11)/4732/13/$6.00

that the actual time course of folding is as realistic as possible. Note that this is not necessary for calculating thermodynamic quantities, since it is only necessary that the moves satisfy detailed balance. In fact, O'Toole and Panagiotopoulos[21] found that using the dynamic moves for the three-dimensional cubic lattice system led to sampling problems. At low temperature they observed a hysteresis for the average energy. This is probably caused by trying to sample below the glass transition. This transition was predicted[10] and explicitly shown[9] to exist in these simple lattice systems. O'Toole and Panagiotopoulos were able to circumvent this problem by using a more sophisticated sampling procedure based on the Rosenbluth and Rosenbluth chain growth algorithm.[22] They studied chains as long as 48 monomers and have used this technique to study thermodynamically significant low energy conformations.[23] Camacho and Thirumalai have used Monte Carlo techniques to examine the relationship between the folding and collapse temperatures and the effect on the folding time.[24]

These previous studies used the standard Monte Carlo technique. The simulation is run at a given temperature and various averages are computed. To obtain thermodynamic quantities for a different temperature, another simulation (at the new temperature) is performed. However, it is possible to extract information about temperatures other than the simulation temperature using a technique often called the *Monte Carlo histogram method*.[25,26] Using this technique one can calculate an approximate density of states for the system, which can be used to calculate any thermodynamic quantity of interest over a range of temperatures. Because of the small system sizes used, we can obtain accurate results over a broad range of temperatures. In particular, we can extrapolate into the glass region where running normal simulations becomes extremely difficult and time consuming. The technique also facilitates the finding of peaks or zeros of various thermodynamic functions. It can be used to calculate extensive quantities like the free energy or entropy of the system, which are difficult to extract by the standard Monte Carlo procedure. One cannot only vary the temperature, but also the various parameters in the potential. One can examine how the system behaves thermodynamically at a range of parameter values without the need to run new simulations. This in-depth analysis of the thermodynamics has enabled us, along with several others, to begin to connect the behavior and properties of these simple model systems with those of real proteins.[7,8]

II. MODEL AND METHODS

A. Definition of model and dynamics

The model studied was a simple three-dimensional cubic lattice polymer. All chains were 27 monomers long with two different monomer types. The potential was a contact interaction between monomers that are nearest neighbors on the lattice (but that are not linked covalently). The energy of a contact was E_l for a pair of the same monomer type and E_u for a pair of different monomers. This model was the same one used in our previous study[9] (which can be consulted for a more detailed explanation of the model). The energy function is

$$E = N_l E_l + N_u E_u, \tag{1}$$

where N_l is the number of contacts between monomers of the same type (like contacts) and N_u the number of contacts between monomers of different types.

Although it is easiest to express the energy function in terms of the E_l and E_u variables, some properties of the model are more clearly understood by considering an equivalent set of parameters, E_{avg} and Δ, defined as follows:

$$E_{\text{avg}} = \tfrac{1}{2}(E_l + E_u), \quad \Delta = (E_u - E_l). \tag{2}$$

E_{avg} represents the overall drive toward forming contacts or compacting the chain. If it is less than zero, contact formation will be favored. Δ determines the heterogeneity of the heteropolymer. In the limit that $\Delta = 0$ the model becomes a homopolymer. In our previous work,[9] $E_{\text{avg}} = -2$ and $\Delta = 2$, giving values -3 and -1 for E_l and E_u respectively.[27] This insures that the chain collapses rapidly, compared to folding, and that the minimum energy state is a maximally compact cube for the sequences considered here. When we say the chain has folded we mean it is in the native state. This is distinguished from collapse, which refers to chains that can be in any compact conformation. The same parameters were used for part of the results presented here. In addition we show how the model behaves as these parameters are varied.

The move set is shown in Fig. 1 and consists of local one- or two-monomer moves. These moves are the standard set used in lattice polymer simulations. They are believed to produce reasonably realistic dynamics (see Refs. 28–31 for details). For thermodynamics calculations, it is not necessary to use a move set with this property, and other move sets have been used.[21,23] There are, however, two potential problems with realistic dynamic move sets: ergodicity and

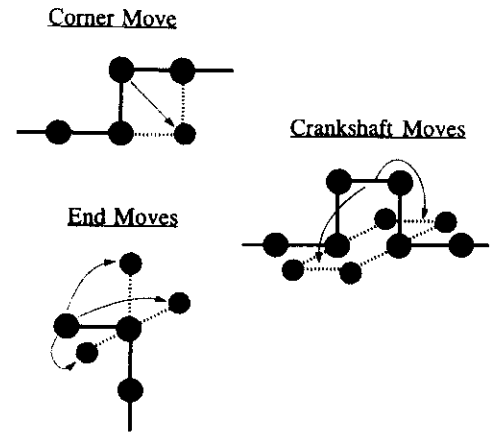


FIG. 1. The three types of moves used in the simulations. The light circles represent the possible lattice points to which a given monomer can move, provided that that point is not occupied. In the case of the end and crankshaft moves, one of the possible moves is picked at random. Note that the corner and crankshaft moves are exclusive: A non-end monomer can only make one or the other depending on the position of its neighbors along the chain.

glassy behavior. If one is interested only in kinetics, then these are not really problems but rather properties of the model. For thermodynamic calculations, inaccessibility is not a problem either, since we can consider the definition of the model to include only the states accessible by the move set specified (hence it will be by definition ergodic). Therefore, as long as the minimum energy cube state is accessible from an unfolded chain, there will be no problems.

The glass transition presents a more difficult problem for thermodynamic calculations. At low temperatures, the dynamics of the system slow down substantially. In particular, the correlation times become quite large and it takes longer to explore conformational space. This gives rise to two errors in the Monte Carlo calculations. First, it takes a long time for the system to relax, making it difficult to get the system into thermal equilibrium. This contributes a systematic error to the results and can give rise to the hysteresis effect seen in the previous studies.[21,32] Second, because of the long autocorrelation times subsequent samples are no longer statistically independent of each other. This has the effect of increasing the variance of any observable (and the corresponding statistical error). The actual variance is given by

$$\sigma^2_{\text{actual}} = \left(1 + 2\frac{\tau_{\text{ac}}}{\tau_{\text{samp}}}\right)\sigma^2, \tag{3}$$

where σ^2 is the usual variance calculated from the samples. τ_{ac} is the integrated autocorrelation time and τ_{samp} is the number of time steps between samples (both measured in units of Monte Carlo steps).[33,34] As τ_{ac} increases, longer simulations are necessary to get statistically reasonable results. That is, in a simulation of N steps there will be N/τ_{ac} "effectively independent samples."[34] At low temperatures the time required to obtain enough independent samples becomes enormous. One solution is to use a different move set, such as the Rosenbluth chain growth algorithm.[22,21,23] The autocorrelation time for this move set does not increase as rapidly, allowing simulations at lower temperatures. Our solution instead is to run the simulations well above the glass transition temperature, and then to use the histogram method to extrapolate to lower temperatures.

B. Techniques for calculating thermodynamics quantities

For short enough chains (usually in two dimensions) it is possible to enumerate all lattice conformations and thereby calculate the partition function for the system along with any other quantity of interest.[13] For the three-dimensional 27 monomer chain, it is not practical[35] to enumerate all conformations but it is possible to enumerate all of the maximally compact (cube) conformations. One could then approximate the partition function by just summing over the cube states. Previous studies have used this method to calculate thermodynamics quantities of this system. It was hoped that at low temperatures this approximation might be reasonable. We show later that, although it can give rough qualitative results, using only the cube states leads to appreciable errors.

Since exact enumeration is impossible, Monte Carlo sampling is used for calculating thermodynamic quantities. The usual technique runs the simulation at a given temperature, collecting samples to determine thermal averages. The process is repeated for several temperatures to get the averages as a function of temperature. There are several drawbacks to the standard Monte Carlo procedure. Calculating extensive variables like the free-energy or entropy is difficult. Also, if one wants to find peaks or zeros of a given quantity (like the specific heat peak, to identify the transition temperature), one must scan over a range of temperatures to located the critical value.

It is possible to extract more information from a single Monte Carlo run than just the thermal averages at the temperature the simulation was performed. The technique is called the histogram method or density of states method. It has a long history and it has been rediscovered recently by a variety of authors.[25,26,36–39] The actual Monte Carlo sampling algorithm itself is unchanged. But instead of just calculating thermal averages, one keeps track of the number of times a specific energy is encountered in the simulation; i.e., an energy histogram is calculated. This histogram, $h(E,T)$, measures the probability of energy E occurring at temperature T. It is equal to the thermal average of the density of states:

$$h(E,T') = \frac{n(E)e^{-E/T'}}{Z(T')}, \tag{4}$$

where $Z(T')$ is the partition function at temperature T'

$$Z(T') = \sum_E n(E)e^{-E/T'}, \tag{5}$$

and $n(E)$ is the density of states for energy E (the number of conformations with energy E). The Boltzmann factor k_B has been set equal to 1 and T' is the temperature of the simulation. One now has the density of states up to a multiplicative factor:

$$n(E) = h(E,T')e^{E/T'}Z(T'), \tag{6}$$

where $Z(T')$ is the unknown multiplicative constant. For intensive quantities, thermal averages are calculated using

$$\langle \mathcal{O} \rangle(T) = \frac{\Sigma_E \mathcal{O}(E)n(E)e^{-E/T}}{\Sigma_E n(E)e^{-E/T}} = \frac{\Sigma_E \mathcal{O}(E)h(E,T')e^{-E/T+E/T'}}{\Sigma_E h(E,T')e^{-E/T+E/T'}}. \tag{7}$$

Note, $Z(T')$ cancels out of the above expression.

If one is interested in calculating extensive quantities like the free energy or the entropy, it becomes necessary to determine the constant $Z(T')$.[40] For our system, it is possible to calculate this constant and therefore obtain the density of states. To determine this constant, all we need is the multiplicity of any energy state. For example, the sequences we study have a nondegenerate ground state. This means $n(E_{\text{gs}}) = 1$, where E_{gs} is the energy of the lowest energy cube. $Z(T')$ can then be determined and the free energy is then calculated using $F = -T \log Z$ and Eq. (5).

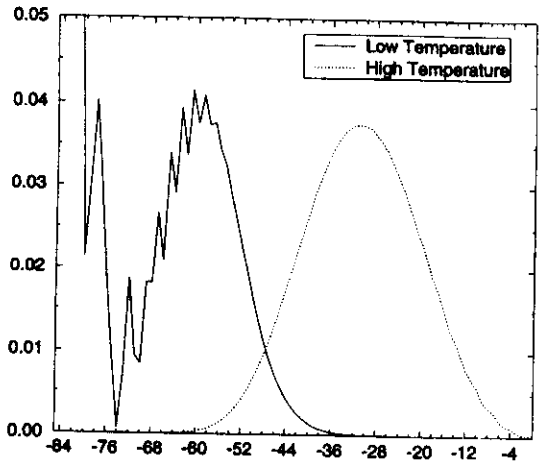


FIG. 2. Energy histograms taken at high ($T=3.15$) and low ($T=1.41$) temperatures. Note that for each histogram there are regions of energy that are not sampled at all. In particular at high temperatures the ground state is not probed while at low temperatures the high energy (unfolded) conformations are not probed.

There is a limit to the temperature range over which Eq. (7) is valid. For temperatures too far from the simulation temperature, the errors in the density of states calculated from Eq. (6) become significant. At any given temperature, the system is only sampling a given region of phase space.[41] For example, Fig. 2 shows energy histograms at high and low temperatures. For the high T simulations, the ground state is never probed, and likewise for the low T some high energy states are never reached. Consequently, the density of states will be incorrect for regions not sampled properly (in fact it equals zero for regions that are never sampled). This limits the temperature range we can extrapolate from any simulation. One needs to monitor the errors in the density of states. A solution to this problem is the multiple histogram method.[26] The idea is to use several different simulations and patch the histograms together. Although there are some subtleties to this technique, it can be powerful.

For the 27 monomer long polymer used here, a single histogram gives adequate results over the range of temperatures of interest. The reason is that the width of the energy histograms in general scale as $1/\sqrt{N}$, where N is the system size. As we shall show shortly, the system is small enough to insure that the histograms are broad and a large region of phase space is sampled at any given temperature.

III. RESULTS AND DISCUSSION

Several Monte Carlo runs were performed over a range of temperatures. Six sequences were used all with a fixed ratio of monomer types (14 to 13). Table I lists the sequences and the energy of their native states. To get some idea of what a typical (i.e., randomly chosen) sequence is and how it compares to the sequences used in this study, we generated over 10 000 random sequences (with a 14:13 ratio of monomer types). Approximately 1 000 sequences had unique ground states. Figure 3 shows a histogram of the ground state energies for these sequences. The distribution is roughly Gaussian. The most probably energy is approximately -76. One sequence examined (sequence 013) has a typical ground state energy for random sequences. We did not generate any of the minimum energy (-84) sequences at random. The two that we used were both designed: one using a Monte Carlo algorithm[42] and the other by mutating a -82 energy sequence. The thermodynamic averages of several quantities (average energy, contacts, native contacts, specific heat, etc.) were calculated at each of the simulation temperatures. In addition, histograms of the number of like versus unlike contacts (N_l and N_u) were calculated. We chose to histogram these variables instead of the energy directly because these histograms can be used to extrapolate not only other temperatures but other parameter values (E_l and E_u).

TABLE I. The various sequences used in this paper. The last four (005, 006, 007, 013) were generated at random. Sequence 002 was optimized by Shakhnovich (Ref. 42). Sequence 004 is a single monomer mutation of 005 ($B_{13}\rightarrow A$). Both 002 and 004 have the lowest energies possible for the potential used and have native states that are completely unfrustrated. E_{min} is the energy of the native states. These same six sequences were studied in our previous work (Ref. 9) which examined the kinetics of folding.

Run	Sequence	E_{min}
002	ABABBBBBABBABABAAABBAAAAAAB	-84
004	AABAABAABBABAAABABBABABABBB	-84
005	AABAABAABBABBAABABBABABABBB	-82
006	AABABBABAABBABAAAABABAABBBB	-80
007	ABBABBABABABAABABABABBBABAA	-80
013	ABBBABBABAABBBAAABBABAABABA	-76

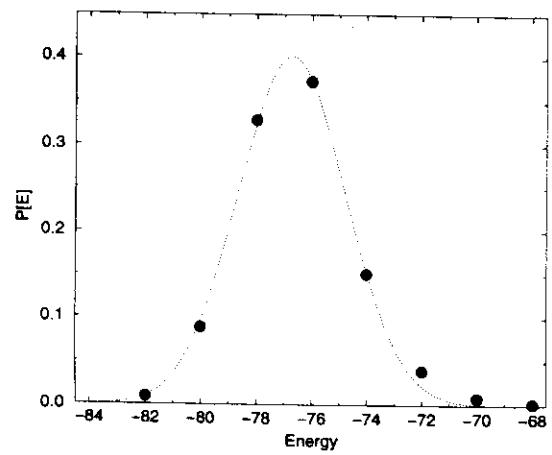


FIG. 3. Ground state energy histogram for nondegenerate sequences. 13 563 sequences with a fix ratio of monomer types (14:13) were generated at random. The energy and multiplicity of the minimum energy cube was determined from exhaustive enumeration of the cube conformations. 1061 sequences (7.8%) were found to have unique ground states. The histogram of ground state energies from these sequences is plotted (solid circles). The dotted line is a least squares fit of a Gaussian to the data. Note, we did not find any minimal energy (-84) sequences in our random sample. The total number of possible sequences is $\binom{27}{14}=20\,058\,300$.

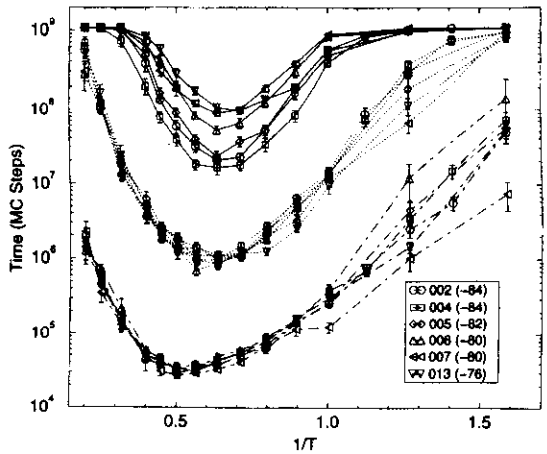


FIG. 4. Folding and collapse times versus inverse temperature. Time is in Monte Carlo steps. The solid lines represent the mean folding time. The middle dotted lines represent the mean compaction time to any cube. The bottom lines represent the mean compaction time to a partially compact conformation with 25 (out of 28) contacts. Error bars are the standard deviation of the mean. Note, simulations were run for a set amount of time τ_{max}. For high and low temperatures some runs were not able to find the folded (or compact) state in this time. In these cases τ_{max} was averaged in, so the times shown are lower bounds to the actual mean first-passage times.

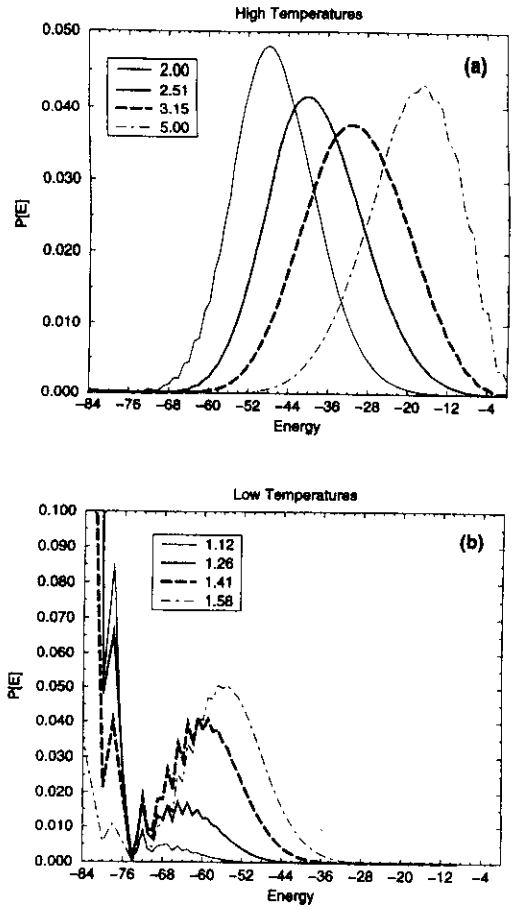


FIG. 5. Histograms of energy for several temperatures. The sequence is number 002. For high temperatures (a) the plots have one peak which moves to lower energies as the temperature decreases. For low temperatures (b) the plots are bimodal.

Figure 4 shows the folding time and collapse times as a function of the inverse temperature. The two different collapse times are the time to find the first cube (i.e., a maximally compact state) and the time to form the first 25 (out of 28) contacts. For high temperatures, the collapse times are sequence independent (*self-averaging*). The folding times are sequence dependent. Similarly, the collapse times below the glass transition are also sequence dependent. The kinetic glass transition temperature was defined in our previous work[9] as the temperature at which the folding time is halfway between its minimum and maximum value (the maximum being determined by the time limit on the simulation and chosen to be much longer than the fastest folding time). Both the folding and collapse time show non-Arrhenius behavior at high temperatures.

A. Histograms: First- vs second-order transitions

Before using the histograms to calculate the density of states, we examined them to determine how much of the phase space is sampled at different temperatures. This (as mentioned above) determines how far we can extrapolate from the simulation temperature using the histogram technique. Figure 5 shows the energy histogram for sequence 002 for several temperatures. Because of the small size of our system, they are all rather broad, with widths of roughly 30 energy units. Examining the behavior of the curves as a function of temperature, we see the both first- and second-order-like behavior of the system. At high temperatures [between 5 and 2, see Fig. 5(a)] a single energy peak moves steadily to lower values. This is what would be expected from a second-order-like transition. At lower temperatures [Fig. 5(b)] the plots now have a bimodal distribution and as the temperature is decreased there is a shift from one peak to the other. This is characteristic of a first-order-like transition. If we examine the histograms as a function of the number of contacts, the same behavior is seen (Fig. 6). At high temperatures the plots are unimodal and shift to a larger number of contacts (higher compactness) as the temperature is lowered. This continues until the maximum reaches roughly 20 contacts around $T=2$. At lower temperatures this peak remains fixed at about 20 contacts and another peak forms at 28 contacts. Because this peak at 28 contacts occurs at low temperatures and when there is a peak at -84 in the energy histograms, we expect that it is due to occupation of the native state and the few other low energy cubes. As the the temperature is decreased there is a shift in population between the two peaks. This is consistent with the idea that there are two thermodynamic transitions: a collapse to compact structures and a folding transition to the native state. The collapse transition occurs at a higher temperature and is second-order-like. The folding transition is first-order-like.

Since the histograms are broad enough we used the single histogram technique in the subsequent calculations. Because we are interested mainly in the properties of the ground state, we chose a temperature which is low enough for the ground state to be sufficiently populated and yet high enough so that we sample as much of the conformation space

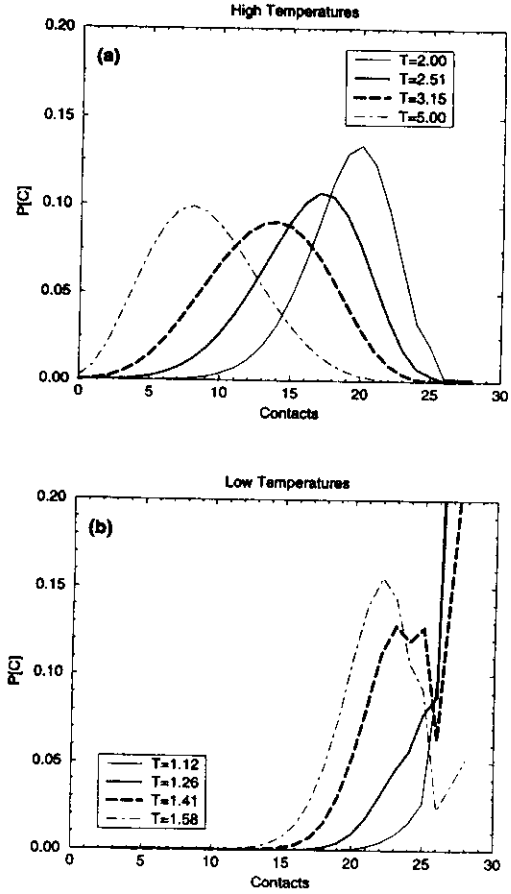


FIG. 6. Histograms of contacts for several temperatures, again for sequence 002. The behavior of the plots is similar to that of the energy histograms (see Fig. 5). As the temperature is lowered a single peak moves from a low to a high number of contacts until it reaches roughly 20 (a). At this point a second peak forms at 28 contacts and we see a first order-like transition between the two (b).

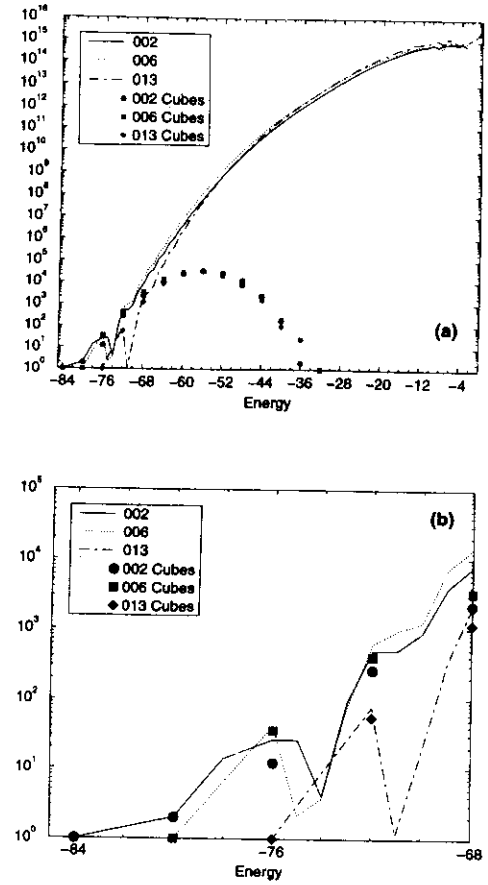


FIG. 7. Density of states versus energy calculated using the Monte Carlo histogram technique. Three different sequences are shown (002, 006, 013) with different ground state energies. The first plot (a) is the full density of states. The lines are from the Monte Carlo calculation. The points show the density of states for just the cube conformations (determined by exact enumeration). The second plot (b) is a blowup showing the low energy region of the first. At high energies the densities of states are sequence independent while at low energies they are strongly sequence dependent.

as possible. At a temperature of 1.58 there is a sizeable peak at the ground state and substantial sampling of the higher energy conformations. Kinetically it turns out that at $T=1.58$ the chains fold most rapidly (i.e., the mean first-passage time to the folded state is smallest). So we expect at this temperature that we are moving most rapidly through the compact conformations. Since this temperature is far enough away from the kinetic glass temperature (which was previously measured to be approximately 1 for this system) we do not have to worry about the problem of long relaxation times which would make it difficult to equilibrate and would require a long sampling time to reduce statistical errors. We calculated the energy autocorrelation function:

$$C_E(t)=\frac{\langle E_\tau E_{\tau+t}\rangle-\langle E\rangle^2}{\langle E^2\rangle-\langle E\rangle^2}, \tag{8}$$

where E_τ is the energy of the system at time step τ. At long times this function should have the following form:

$$C_E(t)\sim e^{-t/\tau_{ac}}, \tag{9}$$

where τ_{ac} is the autocorrelation time.[34] At the temperature $T=1.58$ we get an autocorrelation time of roughly 500 000 Monte Carlo steps. For all our thermodynamic simulations we equilibrated our system for $20\tau_{ac}$ and ran it for 1.08×10^9 steps (2000 times τ_{ac} which gives roughly 2000 independent samples).

B. Density of states

Using the histograms from the Monte Carlo simulations and equation 6 the density of states for the various sequences is calculated. Figure 7 shows the densities for three sequences with low, medium and high values for E_{min} (002, 006, 013). For comparison the cube spectrum, determined by exact enumeration, has also been plotted. At low energies the density of states is different for each sequence. In particular, sequence 002 has more low energy conformations that are not cubes as compared to 006 and 013. All three sequences

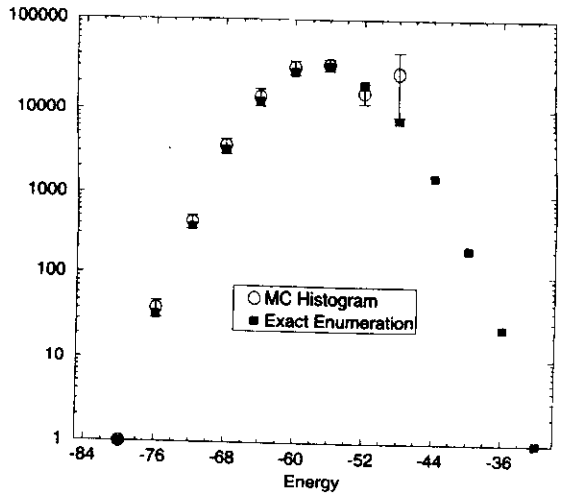


FIG. 8. Comparison of the cube spectrum from exact enumeration and from the Monte Carlo histogram calculation. Error bars are the standard error of the mean from several Monte Carlo runs. The temperature of the simulations was 1.58. At this temperature the average energy is -55.5 and the percent of population in the ground state is 0.2%.

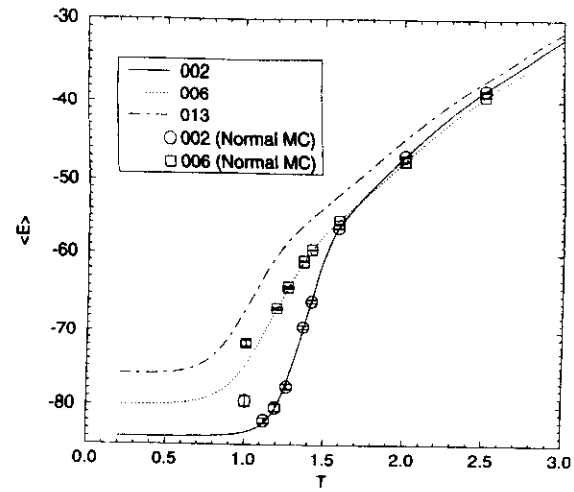


FIG. 9. Average energy versus temperature for three different sequences (002, 006 and 013). The lines are calculated using the histogram technique from simulations at $T=1.58$; the points are from normal Monte Carlo simulations (i.e., they were calculated from the usual averaging technique at several different temperatures). For most temperatures there is excellent agreement between the two. As we approach the glass temperature the normal Monte Carlo technique starts to deviate due to the divergence of the relaxation (equilibration) time of the system.

have a notch in their density, but the gap between the notch and the folded state is larger for sequence 002 than for the others.

At energies below -60 the densities of states for the three sequences are roughly the same. This part of the spectrum is *self-averaging* as expected since it should depend on only the ratio of monomer types. At very high energy there is considerable scatter in the plots. This is due to the poor sampling of this area of conformational space. In particular, the curves in Fig. 7 for sequences 002 and 006 do not even extend to zero energy, indicating that these conformations are not sampled at all. However, we expect that for low temperature thermodynamic calculations this will not pose problems.

As a simple check of the accuracy of the Monte Carlo histogram technique in this system, we compare the exact cube spectrum (from enumeration of all cubes) to the cube spectrum calculated from the histogram data. Remember that there is an unknown normalization factor, which we determined by setting the density of states for the lowest energy cube equal to 1. Figure 8 shows the comparisons. For cubes up to an energy of -52 there is excellent agreement between the histogram calculation and the exact answer. At high energies we see the same sampling problem; cubes with energy greater than -50 are not sampled at all, since they make up a negligible fraction of the conformations at these energies. However, for low temperature calculations the errors should be negligible.

C. Computing thermodynamic quantities

Figure 9 is a plot of the average energy as a function of temperature for the same three sequences (002, 006 and 013) whose densities of states are shown in Fig. 7. At high temperatures ($T>2.5$) all their sequences have roughly the same average energy. At lower temperatures the sequences are no longer self-averaging. The two sequences with the higher energy folded states (006 and 013) have a fairly broad transition while the low energy sequences (002) have a comparatively sharper transition. A similar result is seen in the specific heat, which is plotted in Fig. 10. Sequence 002 has a much sharper and higher specific heat peak which occurs at a higher temperature. The other sequences have broader, smaller peaks. The peak in the specific heat occurs at temperatures slightly higher than the folding temperature (see Table II) and indicates the transition from the unfolded chain

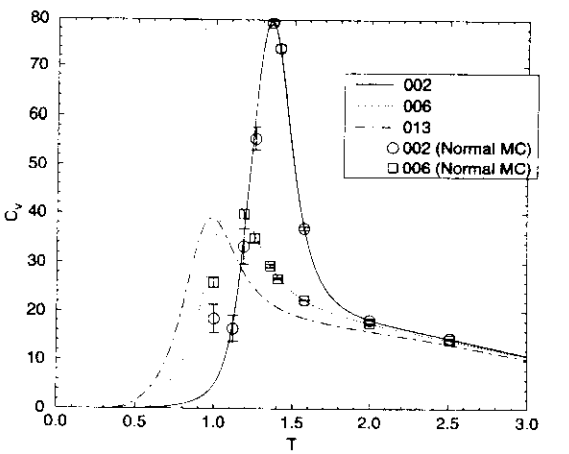


FIG. 10. Specific heat versus temperature for sequences 002, 006 and 013. The lines are calculated using the histogram technique from simulations at $T=1.58$; the points are from normal Monte Carlo simulations (i.e., they were calculated from the usual averaging technique at several temperatures). At the kinetic glass temperature ($T=1$) there is substantial error in the normal Monte Carlo result.

to the collapsed state rather than the transition to the native state. At high temperatures the specific heat is sequence independent.

Also included in Figs. 9 and 10 are data points calculated using the standard Monte Carlo averaging technique for comparison. There is excellent agreement between the histogram curves and the points up until the kinetic glass temperature (which is at $T \approx 1$ for these sequences). At the glass temperature the usual Monte Carlo technique has a problem with the increasing autocorrelation time (all simulations were equilibrated and sampled for the same amount of time). The histogram method, however, allows us to probe beyond the glass temperature since the low energy states can be sampled accurately at higher temperatures (see Fig. 8).

One useful feature of the histogram technique is the ability to determine extrema and zeros of thermodynamic functions. For example, the folding temperature T_f is the temperature at which the population of the native state equals one-half:

$$P_{\text{nat}}(T_f) = \frac{e^{-E_{\text{nat}}/T_f}}{Z} = \frac{1}{2} . \tag{10}$$

Once the density of states has been determined, one can numerically solve for T_f using any standard root-finding algorithm.[43] Figure 11 plots $P_{\text{nat}}(T)$ for the three sequences along with the folding temperatures. Also shown is a plot of the probability of being semicompact (which we define as structures have 20 or more contacts):

$$P_{c_{20}}(T) = \frac{\Sigma_E \Sigma_{C \geq 20} n(E,C) e^{-E/T}}{Z} , \tag{11}$$

where $n(E,C)$ is the density of states as a function of energy and contacts. Note that in order to compute this quantity we need to keep track of histograms as a function of energy and contacts. Histograms of just the energy would not have allowed us to sort out the compact states from the noncompact ones. Similar to T_f the compaction temperature T_c occurs when the probability of being semicompact (20 contacts) equals one-half. Figure 11 shows the $P_{c_{20}}$ curves along with the values for T_c. We chose 20 contacts because that was the point at which the histograms (see Fig. 6) changed from their second-order- to first-order-like behavior. Therefore, we expect that measuring this quantity will probe the first transition from random coil to globule (semicompact) states.

As expected the folding temperature (T_f) is sequence dependent. Also, the transition curves for folding are much sharper than they are for collapse. The lower the energy of the ground state the higher the folding temperature for that sequence. In contrast, the compaction temperature (T_c) is almost sequence independent (it varies by only 4% versus a 43% difference for T_f). One would expect the compaction temperature to be self-averaging since it should depend on the average composition of the sequence (which is the same for all sequences used in this work). At the compaction temperature the native state occupation (P_{nat}) is very small. This is consistent with the previous observation from the histograms (see Figs. 5 and 6). There are two separate thermodynamic transitions: collapse from a random coil and then folding to the native state.

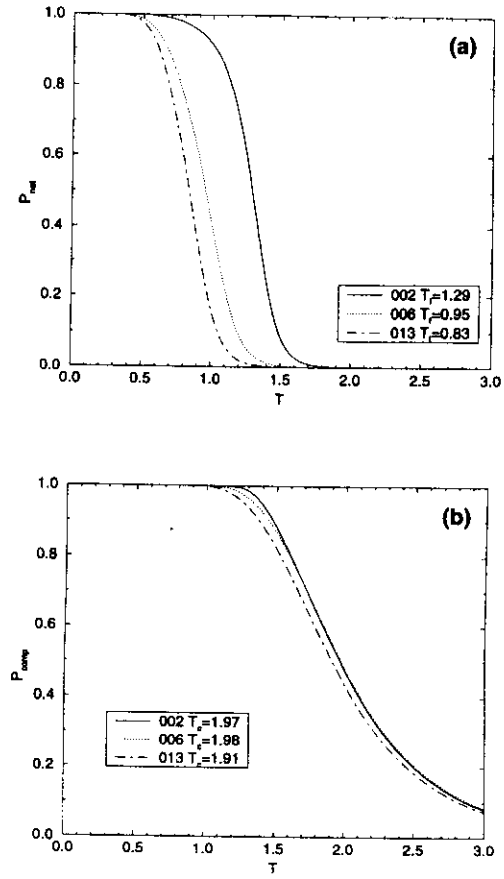


FIG. 11. Folding and collapse transitions for sequences 002, 006 and 013. Figure (a) is a plot of the probability to be in the native state (P_{nat}) versus temperature, while Fig. (b) plots the probability to be compact (20 out of 28 contacts, $P_{c_{20}}$). The folding temperature is defined as the temperature at which $P_{\text{nat}} = \frac{1}{2}$. The collapse temperature is defined similarly. The folding temperature is a sequence-dependent quantity while the collapse temperature is roughly sequence-independent (self-averaging). We expect the collapse transition to depend on the ratio of monomer types (i.e., the overall drive to compactness) and therefore it should not depend on the specific sequence.

It is clear from Fig. 7 that using just the cubes to calculate thermodynamics can lead to potentially large errors. Table II compares the folding temperature calculated using

TABLE II. Comparison of the folding temperature T_f calculated using the full density of states (from the histogram method) and just the cube states (from exact enumeration). Numbers in parentheses are the uncertainty in the last digit. The last column is the percent error of the cube-only calculation.

Run	Full DOS	Cubes	Percent error
002	1.29(2)	1.763	37.2%
004	1.26(1)	1.695	34.1%
005	1.15(2)	1.429	24.2%
006	0.94(6)	1.049	11.5%
013	0.83(5)	0.935	12.6%

the full density of states with the temperature calculated solely on the basis of cube states. The cube results consistently overestimate the folding temperature. This is not surprising since many low energy non-cube states are being neglected. These states will reduce the stability of the native state. The cube approximation is better for the non-folding high energy sequences than for the low energy sequences. Consequently, it fails more seriously for the sequence we are most interested in, namely the good folding sequences.

D. Combining kinetics and thermodynamics—unfolding time

Much of the work (both experimental and theoretical) on protein folding deals with the forward process (unfolded to folded). However, studying the reverse process, the unfolding of nascent proteins, may not only provide a wealth of information, but may also be a great deal easier. In unfolding simulations, the initial condition is well-defined (the folded state) and by varying the various parameters it is fairly easy to induce unfolding. There have been several works examining unfolding using detailed molecular dynamics simulations.[44,45]

Using the data previously calculated, we can compute an unfolding time for our lattice chains. We first make the two-state assumption, namely that there is an unfolded state (U) that is in thermal equilibrium with the folded state (F):

$$F \rightleftharpoons U. \tag{12}$$

We can then calculate an unfolding time (τ_u) as follows:

$$\tau_u(T) = \frac{[F]}{[U]}\,\tau_f(T) \tag{13}$$

where $[U]$, $[F]$ are the populations of the unfolded and folded state respectively and $\tau_f(T)$ is the folding time as a function of temperature. The ratio $[F]/[U]$ is given by $P_{\rm nat}(T)/(1-P_{\rm nat}(T))$, using $P_{\rm nat}(T)$ defined by Eq. (10). Figure 12 shows a plot of the unfolding time versus $1/T$ for several sequences. Unlike the folding time (Fig. 4), the unfolding time has a much simpler behavior. For temperatures above the kinetic glass temperature (T_g), the unfolding times vary almost linearly with $1/T$ and have slightly different slopes for the various sequences. The slopes are roughly proportional to T_f; sequences 004 and 002 have the steepest slopes and 013 has the shallowest. In contrast to the folding time, which shows clear non-Arrhenius behavior over this temperature range, the unfolding time is nearly Arrhenius. We expect due to the large enthalpic barrier that entropic effects are less noticeable in unfolding than folding. At low temperatures the unfolding time rolls off due to the cutoff in the simulation time (this is the same roll-off seen in the folding time at low temperatures; see Fig. 4).

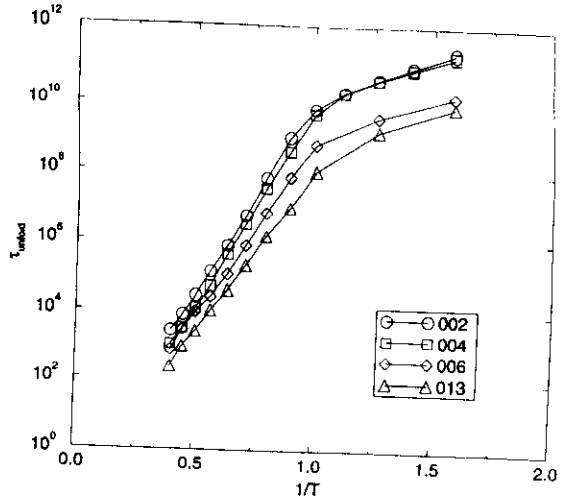


FIG. 12. Unfolding times versus $1/T$ for several sequences. The time is plotted on a log scale. The linear relationship between time and $1/T$ for temperatures greater than T_g is much simpler than the behavior of the folding time.

E. Exploring parameter space

All the results so far have been for simulations with contact energies $E_{\rm avg}=-2$ and $\Delta=2$ ($E_l=-3$ and $E_u=-1$). These values were chosen to insure that the ground state (native state) would be a cube. For sequences like 002 and 004 there exists a cube (out of the 103 346 possibilities) that has no weak contacts (contacts between different monomer types). This cube will be the ground state as long as $E_l=(E_{\rm avg}-\Delta/2)<0$, irrespective of what E_u is (i.e., E_u can be greater than zero). We call these sequences *unfrustrated*. All of the other sequences have at least one weak contact between unlike monomers even in their lowest energy cube (see Fig. 13). These are *frustrated* sequences. For frustrated sequences, it is not clear that the minimum-energy cube will be the minimum-energy conformation. There may be some noncube conformation with fewer total contacts but more good contacts than the cube conformation. For sequence 005, which has 27 good contacts in its minimum cube conformation, and 006 and 007, which have 26 good contacts, there are no other conformations that have more good contacts. Consequently, the cube conformations will be the ground states as long as $E_u=(E_{\rm avg}+\Delta/2)<0$.[46] For sequence 013, which has 24 good contacts in its minimum-energy cube, it is possible that there is some non-cube conformation with more than 24 good contacts. We cannot exhaustively enumerate all the noncube conforma-

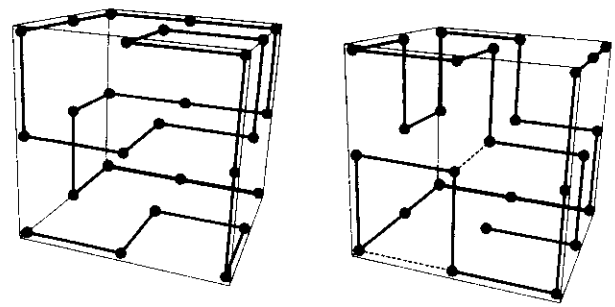


FIG. 13. The native conformations of sequence 002 (left) and 006 (right). Note that 006 has 2 "weak" contacts (indicated in the figure with dotted lines) in its lowest energy state.

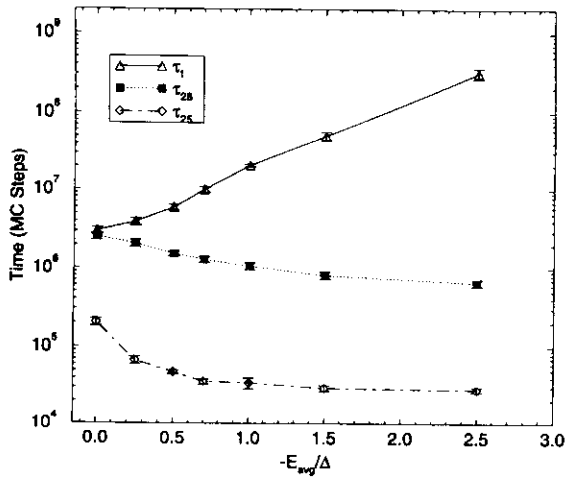


FIG. 14. The minimum folding time and the compaction times plotted as a function of E_{avg}/Δ for sequence 002. The line with the triangles is the mean folding time (i.e., the time to find the native state). The squares are the mean time to find the first cube state (not necessarily the native state) and the diamonds are the mean time to find a conformation with 25 contacts. For each of these times, there is a minimum point in the point of time versus temperature (see Fig. 4). It is that minimum (fastest) time that is plotted here for each value of E_{avg}/Δ.

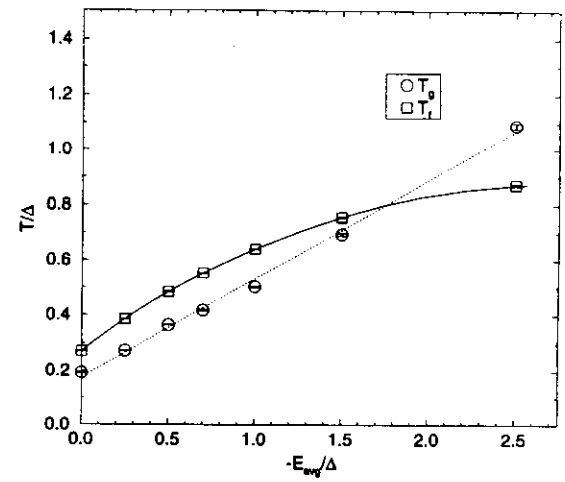


FIG. 15. A plot of the folding temperature (T_f) and the kinetic glass temperature (T_g) as a function of the average drive to compactness (E_{avg}). Both the temperatures and the energies have been scaled by Δ. The dotted line is a linear fit of the glass temperatures. The solid line is not a fit to the T_f points but was calculated explicitly via the histogram technique. For values of E_{avg}/Δ approximately less than 1.7 $T_f > T_g$ so the chains will fold before hitting the glass transition. For E_{avg}/Δ greater than 1.7 the glass transition occurs before folding so the chains do not reach their ground state within the simulation time.

tions but we can state empirically that no conformation with a lower energy was found in any of the Monte Carlo runs (see Ref. 46).

We now explore how the model behaves as we vary the potential parameters. Specifically, how do the various kinetic and thermodynamic properties depend on these parameters? Sequence 002 was examined in detail. As mentioned, this sequence has an unfrustrated cube conformation (see Fig. 13). This cube will be the ground state whenever $E_l < 0$. If E_l is greater than zero then the minimum energy conformation is the completely unfolded chain with no contacts. This clearly would not represent a protein under folding conditions; thus we ignore this region of parameter space. Several simulations with various values[47] of E_{avg}/Δ ranging from 0 to approximately 2.5 were run.[48] Figure 14 shows a plot of the folding and compaction times versus E_{avg}/Δ. The times plotted in these figures are taken from the temperature with the fastest time (i.e., the minimum point in Fig. 4) for each value of E_{avg}/Δ. As the absolute value of E_{avg}/Δ is decreased from 2.5 to 0, both compaction times (the time to form 25 and 28 contacts) increase, although the change is small. This is expected since the drive to form contacts decreases as $|E_{avg}/\Delta|$ is reduced. For the folding time there is an opposite and more dramatic effect. Decreasing $|E_{avg}/\Delta|$ decreases the folding time (τ_f), almost two orders of magnitude. Also, at $E_{avg}=0$, τ_f almost equals τ_{28} (the time to make 28 contacts). What is happening is that as $|E_{avg}/\Delta|$ decreases, we are destabilizing non-native cubes relative to the native state. At large $|E_{avg}/\Delta|$, these low energy non-native cubes behave as traps slowing down the folding rate. Reducing $|E_{avg}/\Delta|$ increases their energy relative to the native cube eliminating them as traps. This in turn increases the folding rate. Alternatively, as $E_{avg}/\Delta \to 0$, E_u increases until it becomes positive. At this point making weak (incorrect) contacts is unfavored relative to breaking them. This keeps the chain from forming these weak, incorrect contacts which would trap it in states different from the native state. The chain is more effectively funneled into the native conformation; i.e., the first cube made is almost always the native one. By varying E_{avg}/Δ we can control the time-scale separation between folding and collapse to maximally compact states.

Next, we examined how the folding temperature (T_f) and the kinetic glass temperature (T_g) varied as a function of E_{avg}/Δ. In particular, for which values of E_{avg}/Δ is T_f greater than T_g? For these values of E_{avg}/Δ, the chain will fold before the native state becomes inaccessible. The histogram method is used to calculate T_f for various values of E_{avg}/Δ. The technique is the same as the one used to extrapolate to different temperatures. In this case one needs histograms as a function of good and weak contacts (which can be calculated from the energy-contact histograms previously used). To calculate T_g we must run simulations at various values of E_{avg}/Δ since there is no way to extrapolate as in the case of T_f. The two temperatures (T_f and T_g) are plotted for sequence 002 in Fig. 15. Note that we plot the temperature normalized by Δ (i.e., in units of Δ) just as we plot E_{avg} normalized by Δ.[48] The glass temperature varies almost linearly with E_{avg}/Δ. Previous work on glass transitions in heteropolymers have shown that the transition occurs after the collapse of the system.[49,50] In fact it was shown that the polymer needs to collapse in order to have a glass transition. This is consistent with the behavior we see: as $|E_{avg}/\Delta|$ increases so does the collapse temperature (T_c). In fact as we will see shortly $T_c > T_g$ for all values of E_{avg}/Δ. As $|E_{avg}/\Delta|$ increases, the depth of the local minima increases relative to the unfolded state. It becomes harder to

escape from local traps so the chain "freezes" at higher temperatures.

The folding temperature also increases as the absolute value of E_{avg}/Δ increases but reaches a limiting value. This behavior of T_f with E_{avg} is easy to understand. As $|E_{avg}/\Delta|$ is increased, the stability gap becomes larger; hence T_f increases. However, it eventually asymptotes to a limiting value of approximately 0.87. This limiting value turns out to be precisely the folding temperature that would be calculated using just the cube conformations. In Table II we see that for sequence 002 $T_f = 1.763$ for the cube-only calculation. In calculation $\Delta = 2$ so we need to divide the temperature by 2, giving 0.88. The reason T_f approaches the cube-only value is that as $|E_{avg}/\Delta|$ increases the cube states decrease in energy more than the noncube states since they have more contacts. Eventually, at low enough E_{avg}/Δ all the low energy states are just cubes. It is the low energy states that determine T_f. Once T_f reaches this limit, it is unchanged by further changes to E_{avg}/Δ since the relative energy between the various cube states is determined by Δ which we are holding constant at 1.

Looking at both the T_f and T_g lines we see there is a key point at which T_g becomes greater than T_f. At $E_{avg}/\Delta \approx 1.7$ the chains become glassy before they become thermodynamically stable; consequently, they will not be able to fold. For E_{avg}/Δ greater than 1.7, $T_g > T_f$. The chain is trapped in local minimum and will not find the native state within the simulation time. When E_{avg}/Δ is less than 1.7, $T_g < T_f$, so the chains fold to the native state and will be thermodynamically stable before the dynamics slow down. By including the results for the collapse transition temperature (the temperature at which half the chains have at least 20 contacts) a qualitative phase diagram can be drawn (see Fig. 16). There are four regions: random coil, collapsed globule, folded and collapsed frozen state. The phase diagram is very similar to other lattice models and theoretical calculations of heteropolymers.[51,52] The vertical dotted line represents the transition from the glassy to folded phase. To the right of it (i.e., large $|E_{avg}/\Delta|$) $T_f/T_g < 1$ and to the left of the line (small $|E_{avg}/\Delta|$) $T_f/T_g > 1$. As $|E_{avg}/\Delta|$ is decreased the folding and collapse curves converge. This is the same behavior observed in the kinetic data (see Fig. 14). As E_{avg}/Δ approaches zero, E_u becomes positive. Collapsed states with weak contacts will be unfavored relative to states with no contacts. This drives the polymer to form only correct (good) contacts, so the chains will collapse almost directly to the correctly folded state. Another way to understand this is that non-native cube conformations are unfavored as E_u increases. At $E_{avg}/\Delta = 0$ roughly half of the cubes for sequence 002 will have positive energies. This removes them as possible kinetic traps. For all values of E_{avg}/Δ, the collapse temperature is greater than not only the folding, but also the glass, temperature. Previous heteropolymer mean-field calculations[49] show that T_g must be less than T_θ, the collapse temperature, consistent with what we find here. Perhaps most interestingly, we see that by modulating a single parameter in our model we obtain a range of qualitatively different folding behaviors. Recently, work has been done on the classification and examination of the various possible folding regimes with a comparison to experimental data.[7]

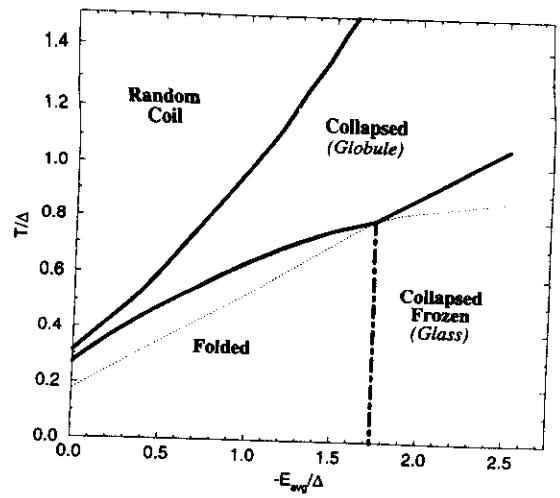


FIG. 16. Phase diagram for sequence 002. There are four regions: random coil, collapsed globule, collapsed frozen, and folded. The solid line between the random coils and the collapsed globule state is the collapse transition, the temperature at which half the chains have 20 contacts. The second solid line is equal to either T_f or T_g, whichever is greater at that value of E_{avg}/Δ. The dotted line equals the lesser of T_g or T_f. The vertical dash-dotted line shows the transition from the folding region (where $T_f > T_g$) to the frozen region. It occurs at $E_{avg}/\Delta \approx -1.7$.

IV. CONCLUSIONS

We have continued our comprehensive analysis of the 27 monomer cube lattice heteropolymer. The Monte Carlo histogram method proved extremely useful, allowing us to determine the density of states for this system and then to calculate a broad range of thermodynamic quantities. In particular, the method overcame the problem of dynamical slowing down at low temperatures. Like many other heteropolymer studies (both analytical and numerical) we find two different transitions: a collapse transition with a roughly sequence-independent collapse temperature and a folding (to the native state) transition with a sequence-dependent folding temperature. The collapse transition has a second-order-like behavior and the folding transition seems first-order-like. The good folding sequences, i.e., the sequences that are stable and have fast folding times, have sharper, more clearly defined transitions, as viewed from the temperature dependence of the average energy and the specific heat. Combining kinetic and thermodynamic data, we studied the unfolding behavior of the system. The unfolding rate has a much simpler temperature dependence than the folding rate. The rate varies roughly linearly with $1/T$ for a broad range of temperature up to the glass point.

After the density of states was obtained we were not only able to extrapolate to different temperatures but also to different parameter values of the energy function. The systems were examined as a function of the average drive toward compactness, where the average drive is the average of

the two contact energies divided (normalized) by their difference (E_{avg}/Δ). There is a specific value for this energy drive at which the kinetic glass temperature became greater than the folding temperature, indicating that the system would no longer be able to fold within the simulation time due to trapping. We constructed a phase diagram as a function of this average drive and temperature (also normalized by the splitting). As the average energy drive is reduced, the two transitions, collapse and folding, converge. At zero-average drive the system collapses almost directly into the native state.

One criticism of these models is that they are too simple to represent real proteins. However, even in this simple model we see a broad and diverse range of behaviors depending on the parameters used. It seems likely that some of the behaviors of real proteins can be explained by some particular set of parameters. More importantly, it may well be the case that different proteins have different folding behaviors. Some proteins may fold extremely rapidly to the native state, literally collapsing into the native state, while other proteins may have a clear separation in time scales between collapse to a compact but non-native ensemble of structures and the rearrangement of the chain to the final native form. In our model we can interpolate between these two regimes by modulating one parameter. In the future, more realistic models that are still simple enough for a through analysis may reveal more about the properties and functions of real proteins.

ACKNOWLEDGMENTS

We would like to gratefully acknowledge the computational assistance of A. Schweitzer, W. Bialek and the NEC Research Institute. We thank K. A. Dill, H. S. Chan, and P. G. Wolynes for interesting and helpful discussions. We also thank J. Song and A. Schwartz for careful reading of and enlightening comments on the manuscript. N.D.S. is a Chancellor's Fellow at UCSD. J.N.O. is a Beckman Young Investigator. This work was funded by the Arnold and Mabel Beckman Foundation and by the National Science Foundation (Grant No. MCB-9316186).

[1] M. Levitt and R. Sharon, Proc. Natl. Acad. Sci. USA **85**, 7557 (1988).
[2] D. J. Tobias, J. E. Mertz, and C. L. Brooks, Biochemistry **30**, 6054 (1991).
[3] V. Daggett and M. Levitt, Proc. Natl. Acad. Sci. USA **89**, 5142 (1992).
[4] M. K. Gilson, T. P. Straatsma, J. A. McCammon, D. R. Ripoll, C. H. Faerman, P. H. Axelsen, and J. L. Sussman, Science **263**, 1276 (1994).
[5] H. S. Chan and K. A. Dill, J. Chem. Phys. **99**, 2116 (1993).
[6] H. S. Chan and K. A. Dill, J. Chem. Phys. **100**, 9238 (1994).
[7] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, PROTEINS: Structure, Function and Genetics **21**, 167 (1995).
[8] J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, and N. D. Socci, Proc. Natl. Acad. Sci. USA **92**, 3626 (1995).
[9] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **101**, 1519 (1994).
[10] J. D. Bryngelson and P. G. Wolynes, Proc. Natl. Acad. Sci. USA **84**, 7524 (1987).
[11] J. D. Bryngelson and P. G. Wolynes, J. Phys. C **93**, 6902 (1989).
[12] P. E. Leopold, M. Montal, and J. N. Onuchic, Proc. Natl. Acad. Sci. USA **89**, 8721 (1992).
[13] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).
[14] H. S. Chan and K. A. Dill, Annu. Rev. Biophys. Biophys. Chem. **20**, 447 (1991).
[15] E. Shakhnovich and A. Gutin, J. Chem. Phys. **93**, 5967 (1990).
[16] A. Šali, E. Shakhnovich, and M. Karplus, J. Mol. Biol. **235**, 1614 (1994).
[17] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. N. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[18] A. Sikorski and J. Skolnick, J. Mol. Biol. **212**, 819 (1990).
[19] J. Skolnick and A. Kolinski, J. Mol. Biol. **212**, 787 (1990).
[20] J. Skolnick and A. Kolinski, J. Mol. Biol. **221**, 499 (1991).
[21] E. M. O'Toole and A. Z. Panagiotopoulos, J. Chem. Phys. **97**, 8644 (1992).
[22] M. N. Rosenbluth and A. W. Rosenbluth, J. Chem. Phys. **23**, 356 (1955).
[23] E. M. O'Toole and A. Z. Panagiotopoulos, J. Chem. Phys. **98**, 3185 (1993).
[24] C. J. Camacho and D. Thirumalai, Proc. Natl. Acad. Sci. USA **90**, 6369 (1993).
[25] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **61**, 2635 (1988).
[26] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **63**, 1195 (1989).
[27] The units of temperature and energy are such that $k_b = 1$. This still leaves an arbitrary scale factor since the only important quantity is the ratio of energy to temperature, i.e., we could have picked -200 and 200 for E_{avg} and Δ, multiplied all temperatures by 100 and the results would be the same. We have chosen small, integer values for convenience.
[28] P. H. Verdier and W. H. Stockmayer, J. Chem. Phys. **36**, 227 (1962).
[29] H. J. Hilhorst and J. M. Deutch, J. Chem. Phys. **63**, 5153 (1975).
[30] K. Kremer and K. Binder, Comp. Phys. Rept. **7**, 259 (1988).
[31] M. T. Gurler, C. C. Crabb, D. M. Dahlin, and J. Kovac, Macromolecules **16**, 398 (1983).
[32] Note, in their study (Ref. 21) the authors do not attribute the problems with local Monte Carlo moves to a glass transition in the system. However, even though their potential is slightly different than the one used in our study, we feel that their system may also undergoes a glass transition which would explain the hysteresis seen.
[33] H. Müller-Krumbhaar and K. Binder, J. Stat. Phys. **8**, 1 (1973).
[34] N. Madras and A. D. Sokal, J. Stat. Phys. **50**, 109 (1988).
[35] Assuming that the number of conformations scaled like 3^n (a very conservative underestimate) and that one could enumerate a million conformations every second (a somewhat generous overestimate of computational power) then it would take 29 days to enumerate all conformations. If the number scales like 4^n it would take 142 years.
[36] N. A. Alves, B. A. Berg, and R. Villanova, Phys. Rev. B **41**, 383 (1990).
[37] S. Huang, K. J. M. Moriarty, E. Myers, and J. Potvin, Z. Phys. C **50**, 221 (1991).
[38] D. Bouzida, S. Kumar, and R. H. Swendsen, Phys. Rev. A **45**, 8894 (1992).
[39] S. Humar, D. Bouzida, R. H. Swendsen, P. A. Kollman, and J. M. Rosenberg, J. Comput. Chem. **13**, 1011 (1992).
[40] Actually if one is interested in calculating differences of the free energy then it is not necessary to determine this constant since it will only change the free energy by an additive constant. To see this note that $F = -T \log Z$, so if we know $c^* n(E)$ (where $c = 1/Z(T')$ is the constant we do not know) we can get $c^* Z$ which gives: $F = -T \log Z + T \log c$.
[41] This is a finite time problem (i.e., we must run the simulations of a finite amount of computer time). In the infinite time limit, for a finite system, all regions of phase space would be explored, so the histograms from one temperature could be used to extrapolate to all.
[42] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. USA **90**, 7195 (1993).
[43] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes* (Cambridge University Press, New York, 1986).
[44] J. D. Hirst and C. L. Brooks, J. Mol. Biol. **243**, 173 (1994).
[45] A. Li and V. Daggett, Proc. Natl. Acad. Sci. USA **91**, 10430 (1994).
[46] It is a property of the cubic lattice that there are no conformations of 27 length polymers that have 27 contacts. Therefore for sequences 005, 006, 007, that have 27 or 26 good contacts for some cube, there can be no non-cube conformation with more good contacts. (We know there are no cube conformations with 28 good contacts since we can enumerate all the cubes.) For sequence 013 which only has 24 good contacts it is possible that there is a conformation with 26 good contacts which would give it an energy of -78. However, we have found no such state in the Monte Carlo simulations.
[47] Actually for technical reasons we ran the various simulation changing the value of E_u, leaving the value of E_l fixed at -3. However, data were scaled to unit Δ when plotted by normalizing both the temperature and E_{avg}. E_u was varied from -2 to 3.
[48] As mention previously (in Ref. 27) there is an arbitrary scale factor in our unit of energy and temperature. Therefore there are only two independent

parameters in our model from the set of three (T, E_{avg}, and Δ). We choose to vary T and E_{avg}, and will normalize them to unit Δ.

[49] J. D. Bryngelson and P. G. Wolynes, Biopolymers **30**, 177 (1990).

[50] In their work Bryngelson and Wolynes (see Ref. 49) use a slightly different notation and terminology. We discuss the glass transition temperature, T_g they use the term freezing transition and denote the temperature as T_f. This should not be confused what we call the folding temperature (denoted here also as T_f). Also the two transition are different. They calculate a thermodynamic freezing transition while we compute a kinetic transition related to the slow down of the dynamics of the system. See Ref. 9 for details.

[51] E. I. Shakhnovich and A. M. Gutin, Biophys. Chem. **34**, 187 (1989).

[52] A. Dinner, A. Šali, M. Karplus, and E. Shakhnovich, J. Chem. Phys. **101**, 1444 (1994).

Toward an outline of the topography of a realistic protein-folding funnel

J. N. ONUCHIC*, P. G. WOLYNES†, Z. LUTHEY-SCHULTEN†, AND N. D. SOCCI*

†School of Chemical Sciences, University of Illinois, Urbana, IL 61801; and *Department of Physics-0319, University of California, San Diego, La Jolla, CA 92093-0319

Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 3626–3630, April 1995
Biophysics

Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 3626–3630, April 1995
Biophysics

Toward an outline of the topography of a realistic protein-folding funnel

J. N. ONUCHIC*, P. G. WOLYNES†, Z. LUTHEY-SCHULTEN†, AND N. D. SOCCI*

†School of Chemical Sciences, University of Illinois, Urbana, IL 61801; and *Department of Physics-0319, University of California, San Diego, La Jolla, CA 92093-0319

Contributed by P. G. Wolynes, December 28, 1994

ABSTRACT **Experimental information on the structure and dynamics of molten globules gives estimates for the energy landscape's characteristics for folding highly helical proteins, when supplemented by a theory of the helix-coil transition in collapsed heteropolymers. A law of corresponding states relating simulations on small lattice models to real proteins possessing many more degrees of freedom results. This correspondence reveals parallels between "minimalist" lattice results and recent experimental results for the degree of native character of the folding transition state and molten globule and also pinpoints the needs of further experiments.**

Recently a framework for understanding biomolecular self-organization using a statistical characterization of the free-energy landscape of protein molecules has emerged (1–5). Based on the physics of mesoscopic, disordered systems, it can capitalize on the ability to simulate "minimalist" models of proteins, to characterize the folding mechanism through a few energetic and entropic parameters describing the free-energy surface globally. The energy landscape of a foldable protein resembles a many-dimensional funnel with a free-energy gradient toward the native structure. The funnel is also rough, giving rise to local minima, which can act as traps during folding. Most random heteropolymers have numerous funnels to globally different low-energy states just as do glasses and spin glasses. The search through the energy minima of a rough landscape is slow and becomes more difficult as the glass transition is approached. Typically a random heteropolymer will not fold to its lowest free-energy minimum in times less than that needed to explore completely the configuration space if there were no barriers. This supposed difficulty for a natural protein has been called the Levinthal paradox (5). For most random heteropolymers, the search problem of the Levinthal paradox is real, but the guiding forces engineered by molecular evolution can overcome the Levinthal paradox provided they are strong enough, in accordance with the "principle of minimal frustration" (1–4). Most simply, the landscape of a protein funnel is characterized by three parameters: the mean square interaction energy fluctuations, ΔE^2, measuring ruggedness; a gradient toward the folded state, δE_c; and an effective configurational entropy, S_1, describing the search problem size. Our goal here is to use experiments, theory, and simulations to estimate these topographic parameters that determine the folding mechanism. Bryngelson *et al.* (5) classify several regimes of folding. In part of the protein's phase diagram, folding is entirely downhill in a free-energy sense; i.e., as the ensemble of intermediate structures becomes progressively more native-like, the energy gradient completely overcomes the entropy loss. This occurs for folding funnels with very large δE_c and is called type **0** folding. Under thermodynamic conditions near the folding transition midpoint, entropy and energy do not completely compensate each other; thus, intermediates are not present at equilibrium (i.e., a free thermodynamic energy barrier intercedes). In a type **I** transition, activation to an ensemble of states near the top of this free-energy barrier is the rate-determining step. Type **I** transitions occur when the energy landscape is uniformly smooth. When the landscape is sufficiently rugged, in addition to surmounting the thermodynamic activation barrier, at an intermediate degree of folding, unguided search again becomes the dominant mechanism. At this point, a local glass transition has occurred within the folding funnel. If the glass transition occurs after the main thermodynamic barrier, the mechanism is classified as type **IIa.** Here specific kinetic intermediates occur late in folding but are native-like. On the other hand, if the glass transition occurs before the main thermodynamic barrier, intermediates are misfolded traps, which can be described by a multistep chemical kinetic scheme. The details of this type **IIb** mechanism are very sensitive to the thermodynamic state, interaction potentials, and the specific sequence.

Starting with Levitt and Warshel (6), a variety of simple models of protein folding called minimalist have been developed. Recent studies with continuum models by Honeycutt and Thirumalai (7) and others have been interpreted using energy landscape ideas. Another class of minimalist models studies folding of heteropolymers on a lattice using Monte Carlo kinetics (4, 8–11). Both these studies and the exact enumeration schemes pioneered by Dill and co-workers (12) provide a characterization of the energy landscape for such minimalist models. The simplicity of these models strikes some as being terribly unrealistic, since real proteins possess many details not present in most minimalist models such as hydrogen-bonded secondary structure and side chain conformational degrees of freedom important for packing. Each of these has a different energy scale. Can these features be at all taken into account when making the connection between minimalist models and experiments on real proteins without studying highly complex models?

The energy landscape philosophy and the analogy to phase transitions provide the key. The broad mechanism of phase transitions depends only on gross features of the energy function. When appropriately scaled, the part of the phase diagrams relevant to boiling of liquids as disparate as water and xenon can be superimposed. At the empirical level, this remarkable similarity is known as the law of corresponding states (13), and it is also the basic idea of the renormalization group (14).

The energy landscape picture suggests that there is also a law of corresponding states mapping the phase diagram and kinetic mechanisms of real proteins onto those for minimalist models. If separate phase transitions for ordering the additional degrees of freedom possessed by real proteins intervene during folding, a multistep mechanism can still result, but in a major part of the more complex phase diagram, the effect of the extra degrees of freedom in real proteins will be to "renormalize" energy and entropy scales for the protein-folding funnel.

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.

Here we explore a correspondence of real proteins with minimalist protein-folding models that uses an analytic theory of helix-coil transitions in collapsed heteropolymers to effectively renormalize out secondary structure formation. When combined with experimental measurements of the amount of secondary structure, the theory quantifies the effective number of degrees of freedom of a helical protein through the configurational entropy. Dynamical measurements on the molten globule state crudely characterize the energy landscape ruggedness. The funnel's slope is then inferred using the thermodynamics of the molten globule to folded transition.

The reduction of configuration entropy through helix formation in the collapsed state yields an energy landscape comparable in extent or complexity with that for minimalist models with fewer residues but lacking explicitly secondary structure. The corresponding energetic topography for an optimized three-letter code minimalist lattice model roughly corresponds with the energy gradient and ruggedness of a realistic folding funnel.

The parallel between the gross features of the landscape of real proteins and the three-letter code lattice models allows us to quantify aspects of the folding mechanism in real proteins by using computer simulations of the model. By simulating many folding trajectories and characterizing the free energy as a function of several order parameters, we can identify the location of the relevant thermodynamic free-energy barrier, which is rather small, and determine the position of glass transition within the folding funnel. While the broad transition state occurs early, the glass transitions occur rather late in the folding processes for this model. At the denaturation midpoint, folding occurs via a type **IIa** scenario but is rather close to the downhill type **0b** scenario. The details of a protein's folding after the glass transition late in the funnel cannot be studied using the corresponding states principle, but the earlier events can.

Establishing the Correspondence Between Minimalist Models and Real Proteins

Collapsed states have been established as rather general intermediates in folding (15). Some compact intermediates contain a substantial percentage of helical secondary structure. At least two views of the collapsed states are prevalent. Some argue that the molten globule state has a specifically defined tertiary structure comparable to the native protein. Others view the equilibrium collapsed state as still conformationally fluid in terms of the backbone structure resembling a polymer below its θ point (16). Both Kallenbach and co-workers (17) and Engelman and co-workers (18) have found compact states with varying degrees of helical structure, thus suggesting its lability. The two pictures may not be so clearly separable since the guiding forces of the funnel do induce a significant amount of fluctuating tertiary order in the disordered globule. The relevant collapsed states are the dynamic ones of the early stages of folding and not necessarily the equilibrium states found elsewhere in the phase diagram (e.g., the "acid molten globule").

The degree of helical content of equilibrium collapsed states has often been measured to be quite high. The relation between helicity and the conformational entropy shown in Fig. 1 can be found from the theory developed by Luthey-Schulten *et al.* (19), since both depend parametrically on the effective hydrogen bond energy. The theory due to Bascle *et al.* (20) could also be used with appropriate modification. Taking 65% helicity as a reasonable estimate, our theory gives a conformational entropy of $0.6k_B$ per monomer unit. The effective number of states to be searched is related to this Levinthal entropy, e^{S_L/k_B}. Though the states differ in character, the number of states in the mechanism here is comparable to that for the framework model (ref. 21 and references therein; ref. 22). Diffusion–collision calculations assume only the correct helices can be formed and direct construction of the fold from the high entropy random coil. A free chain has an entropy of $2.3k_B$ per monomer unit (23). Unlike the framework picture, the dramatic reduction in entropy upon collapse arises from confinement, indiscriminate helix formation, and the constrained orientation of the helices, not from strong local biases toward correct secondary structure. The molten globule from which further guided searches take place in our mechanism is a collapsed liquid-crystalline polymer. The compact configurations of simple lattice model polymers have an entropy of $\approx 1.0k_B$ per unit considering only the reasonably compact states after fast collapse. Thus the renormalized entropy or scope of configuration space of a 60-amino-acid helical protein is a bit bigger than the 27-mer lattice model often studied.

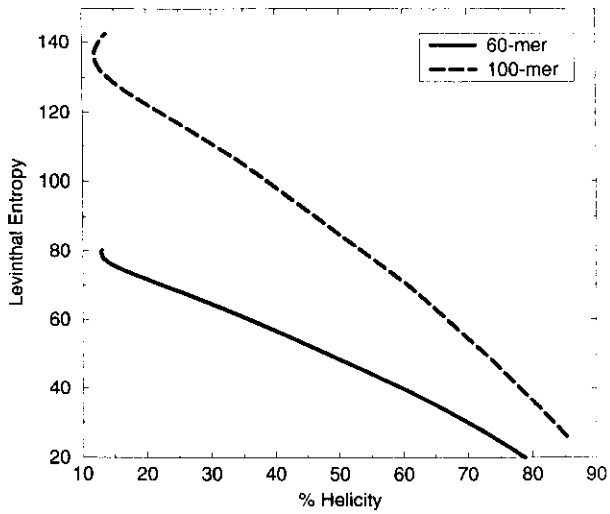


FIG. 1. Configurational (Levinthal) entropy versus helicity according to the theory of Luthey-Schulten *et al.* (19). Both quantities depend parametrically on the effective hydrogen bond strength divided by k_BT.

Comparing the dynamics of free and collapsed chains yields the ruggedness of the landscape. In a free chain, flickering secondary structural elements reconfigure in roughly $\tau_0 = 1$ nsec (24). This time is similar to interdiffusion times over distances of the size of the molten globule diameter as calculated using the Rouse-Zimm theory (25). Thus, to first order, we can be agnostic as to the nature of the underlying move set in comparing real proteins and minimalist models. The dynamics of a condensed molten globule is slower than that for free chains because of transient trapping in low-energy states. The reconfiguration time τ_{reconfig} in a rough energy landscape (5, 26) is given by $\tau_{\text{reconfig}} = \tau_0 \exp(\Delta E^2/2T^2)$.

Few experiments directly measure reconfiguration times within the globule. For lactalbumin, Baum *et al.* (27) observe field-dependent broadening of ^{1}H NMR resonances, suggesting reconfiguration rates slower than 1 per millisecond. Wand and co-workers (28) interpret their NMR studies on the apocytochrome b_{562} molten globule with similar times. In the fastest folding, the downhill scenario (type **0**), folding takes only a few times the typical reconfiguration time. Thus an upper bound on the ruggedness is known since collapsed states can completely fold in times ranging from a millisecond to a second. These estimates for τ_{reconfig} suggest the ruggedness of the energy landscape, at the folding temperature, $\Delta E^2/2T_f^2$, ranges from 11 to 18. The typical size of hydrophobic forces needed for protein collapse gives directly a similar estimate (29). The actual entropy of the molten globule state is lower than S_L, since low-energy states are preferentially occupied. For the 60-amino-acid chain at 60% helicity, the random energy model gives an entropy $S(T) = S_L - \Delta E^2/2T^2 = 21k_B$ to $28k_B$. At the folding transition, the energy loss in falling

down the funnel must equal the temperature times the entropy loss. Thus the stability gap or energy gradient of the funnel is $\delta E_s/T_f = S_L + \Delta E^2/2T_f^2$. The stability gap δE_s measures the difference in energy of the native protein and the average compact state. The dimensionless ratio of the energy gradient of the funnel to the overall ruggedness is then $\{\delta E_s/T_f\}/\{\sqrt{\Delta E^2/2T_f^2}\} \approx 14$. Using the configurational entropy estimate, the thermodynamic glass transition temperature is $T_g = \Delta E/\sqrt{2S_L} \approx 0.6T_f$. If interaction strengths were temperature independent, the thermodynamic glass transition temperature for a compact denatured state would be 160 K. The dynamical glass transition in folded myoglobin actually depends strongly on solvent and occurs in glycerol at 180 K (30). The coincidence might support the ideas of Frauenfelder and Wolynes (31) and of Honeycutt and Thirumalai (7) that some taxonomic substates of folded proteins correspond with the final protein folding intermediates.

Comparing the estimate of gradient-to-ruggedness ratio with the 27-mer simulations shows that the landscape is smoother than landscapes generated for optimally designed sequences using a two-letter code. The ratio between T_f and the kinetic T_g (relevant to the folding time scale) is about 1.3 for the two-letter code sequences (32). The thermodynamic glass temperature for the collapsed states of the two-letter code 27-mers calculated using the random energy model estimate is close to this kinetic T_g. Two-letter code lattice models in the bulk limit usually exhibit ground-state degeneracy, probably connected with microphase separation (33). Yue *et al.* (34) suggest that designing foldable two-letter code proteins is nontrivial.

Simulations have been performed with three-letter codes—i.e., strong interactions for residues of the same kind and weak interactions for different. When the values of the couplings are the same as for the two-letter code, an optimized three-letter code folded configuration still has only correct strong contacts, but most three-letter code compact configurations have fewer wrong strong contacts than for the two-letter code. Optimized three-letter code proteins have a larger T_f and a smaller kinetic T_g than two-letter ones. The T_f/T_g ratio increases to 1.6, close to the ratio for realistic folding funnels. The mechanisms of folding of the two- and three-letter code results differ since the two-letter code model is closer to its global glass transition.

The Folding Scenario for a Realistic Folding Funnel

The corresponding state analysis allows us to sketch the following folding scenario, based on the three-letter code simulations, and to picture a folding funnel whose main features are shown in Fig. 2. We have rendered this folding funnel reasonably accurately to scale. The width is a measure of the entropy, whereas the depth is illustrated with both an energy and two correlated structural scales. Although no one-dimensional scale reflects properly the multidimensionality of the funnel and the multiple minima, the barrier heights in the figure represent ΔE, whereas the total depth is scaled to the energy of the folded state. The molten globule region, a transition state region representing an ensemble of structures that acts as a bottleneck, and a *locally* glassy region are identified based on detailed examination of many folding trajectories for the three-letter code model coupled with numerical measurements of density of states, free energies, and related quantities. Defining several collective coordinates compresses much information about the trajectories into a simple form. Our characterization of the transition state region for a realistic folding funnel differs from the results of Šali *et al.* (35), which apparently model proteins near the border of kinetic foldability.

Two coordinates examined are ones like that of Bryngelson and Wolynes (26), the fraction of angles in their native configuration, A, and the fraction of native contacts, Q. The

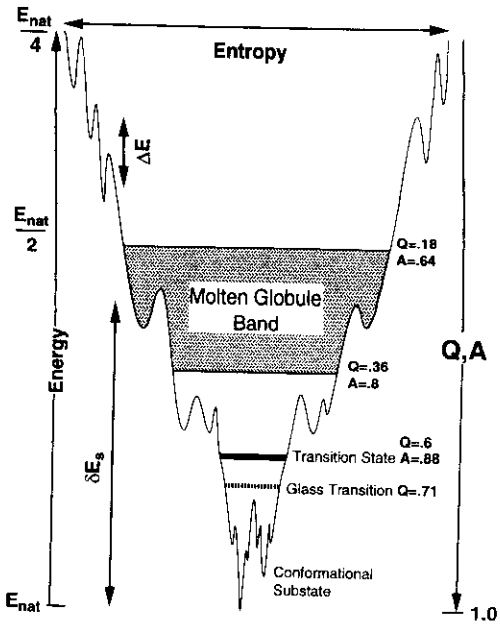


FIG. 2. The schematic funnel for a realistic 60 amino acid helical protein corresponding to the three-letter code. This shows the position of the molten globule, the transition state ensemble, and the local glass transition where discrete trapping states emerge as a function of the order parameters described in the text, the energy E, the fraction of native contacts Q, and the fraction of angles in their native configurations A. Q and A have been normalized to their maximal values for a 27-mer lattice model, 28 and 25, respectively. For the three-letter code, the molten globule is stabilized by nearly half the native energy (E_{nat}) relative to the random coil. The stability gap δE_s quantifies the specificity of the native contacts.

reaction coordinate A only changes by a small amount on each elementary step, so local gradients of free energies are meaningful. For a random coil, the value of A should be ≈ 0.68. Q is intimately connected with the interaction energy function and is useful in describing overall topology. Care must be used in interpreting gradients of free energies with respect to Q since the elementary moves can lead to large Q changes. Both coordinates refer to overall structure features. For larger heteropolymers, additional coordinates describing distinct parts of the chain are needed to define critical nuclei (36, 37) for large single-domain proteins or independent folding of parts of multidomain proteins.

Time series of the reaction coordinates and interaction energy are two-state-like with fast transitions between the folded and unfolded regions. The mean folding time for this sequence of $\approx 3 \times 10^6$ time steps corresponds to 3 msec of real time using the estimates for τ_0. As for simple reactions, the rate for transition between the two main regions depends on short time events.

The duration of a transition event fluctuates, but most events are over in less than 10,000–50,000 time steps. Q and A vary in a correlated manner through the transition. Sometimes they quickly traverse between the stable regions, whereas in other cases they are transiently trapped during the crossing. The duration of the trapping events is much shorter than the average folding time, indicating two-state kinetics.

The Monte Carlo histogram technique (38) was used to determine the density of states as a function of energy and the order parameters. This technique is similar to that used by Hansmann and Okamoto (39) and Hao and Scheraga (40) to determine overall thermodynamics. The densities of states yield the free-energy plots, precisely locate the folding temperature, and determine a local thermodynamic glass transition region where discrete intermediates appear.

At T_f, the free-energy function for the optimal three-letter code 27-mer is plotted as a function of Q and A. Projected onto these plots are two illustrative trajectories out of 86 examined (Fig. 3). The free-energy function is bistable. The disordered globule has $Q \approx 0.28$ and $A \approx 0.73$. The globule has a much greater amount of native structure (Q) than that expected for a random coil, but clearly an enrichment of pair contacts by a factor of 10 does not by itself imply a "unique" structure for a molten globule (41). The two-dimensional free energy has a saddle at $A \approx 0.88$ and $Q \approx 0.6$. The Q value of 0.6 means that each native contact is made three-fifths of the time in an ensemble of configurations of the transition state. This is in harmony with recent observations on chymotrypsin inhibitor folding where the transfer coefficient for mutations at each site ϕ varies between 0.3 and 0.7 (42). Since chymotrypsin inhibitor has a good deal of β-sheet as well as a helix, the agreement may be fortuitous. The superimposed trajectories agree with assigning a transition state region encompassing Q values from $\approx$0.57 to $\approx$0.64 and A values from $\approx$0.84 to $\approx$0.92. Late barriers depending on Q alone are kinetically meaningless, since reactive trajectories jump across such barriers in the Q direction through crankshaft moves in which an entire arm of the protein is retracted into its native position. The A coordinate on the other hand varies only by one or two units per elementary move and is a more appropriate reaction coordinate (1). Because of the flatness of the free energy, the thermodynamic barrier from the free-energy plot is small but broad. There are numerous recrossings of the transition state region caused by the trapping due to the landscape's ruggedness. Thus, as Bryngelson and Wolynes (26) suggest, folding times must be computed using a diffusive picture instead of standard transition state theory, which neglects recrossings.

Monitoring correlated fluctuations of the collected coordinates gives the diffusion constants for Q in the molten globule $D \approx 3.5 \times 10^{-4}$ (correct contacts)^2 per time step. A crude diffusive rate theory that assumes the free energy well is harmonic and that the barrier top curvature equals the well's gives a folding time $\tau_F = 2\pi\tau_{corr}\exp\{F^*/k_BT\}$, where F^* is the activation barrier of $2.4k_BT_f$ from the two-dimensional plot and τ_{corr} is the correlation time for the harmonic fluctuations. τ_{corr} for both A and Q is approximately 20,000 time steps. The resulting $\tau_F \approx 1.4 \times 10^6$ is a bit shorter than the simulated value. Since this system becomes glassy only after the transition region is traversed, landscape ruggedness should be well accounted for by the diffusion picture. The folding time from the Bryngelson–Wolynes (26) approximation is good, but there are more near-ballistic trajectories through the transition region than expected, suggesting the relevance of the frequency dependence of the structural diffusion or of additional geometrical variables.

After leaving the transition region, the protein progresses to become more native-like. Occasionally, the trajectory becomes caught in a few longer lived native-like states whose lifetime is shorter than the average folding time. These discrete states arise from a local glass transition, which can be located by computing $Y = \Sigma_i P_i^2$, where P_i is the Boltzmann occupation of a microstate. Y measures the inverse number of the thermally accessible states and reveals the replica symmetry breaking of spin glasses (43) and of random heteropolymers (44). To define the local glass transition, we compute $Y(Q)$ using only states with a given value of the coordinate Q. Since the protein is of finite size, $Y(Q)$ never vanishes but instead varies from the inverse density of states at Q up to unity. At T_f a rapid rise of $Y(Q)$ occurs at $Q \approx 0.7$, defining a local glass transition (Fig. 4).

At T_f the transition state region occurs before the local glass transition, so folding conforms to a type **IIa** scenario. Kinetic constraints, which vary from sequence to sequence, are encountered after the transition state is reached for strong folders just as in simulations of Honeycutt and Thirumalai (7) and of Chan and Dill and co-workers (10, 12).

The small size of the thermodynamic barrier as opposed to kinetic barriers from transient trapping suggests that proteins are not just overall marginally stable but that a realistic folding funnel describes a marginally stable system even for intermediate degrees of order—surprisingly much like a system near a critical point. *In vivo*, proteins are not poised at T_f but are stable by several k_BT. The additional slope to the funnel's energy gradient should suffice to make folding occur by a downhill type **Ob** scenario where the only intermediates are near native kinetic traps. Since folding does not dramatically speed up with increasing stability once a downhill scenario is reached, perhaps there is no evolutionary drive to greater stability. The combination of marginal stability and proximity

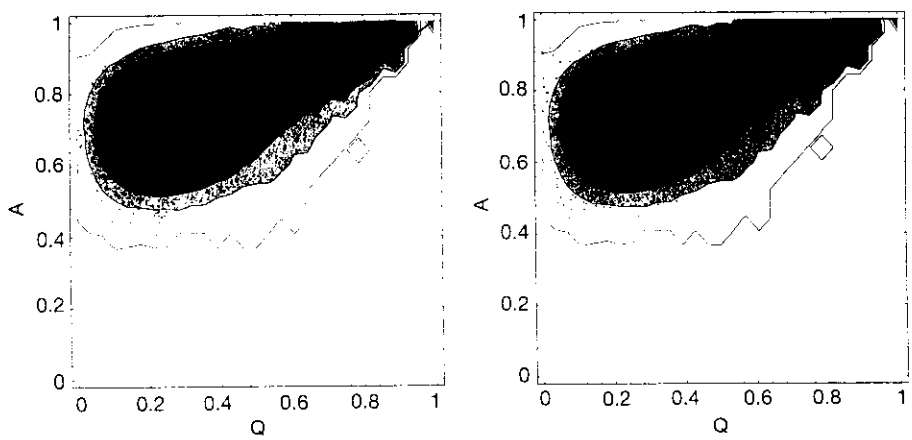


FIG. 3. Two transition trajectories projected onto the Q–A plane. The time span is roughly 25% of the folding time, which is $\approx 3 \times 10^6$ Monte Carlo steps. (*Left*) The transition event occurs in $\approx 10^5$ Monte Carlo steps. For this trajectory, there is some trapping in the transition region. In the early part of the trajectories, the individual points are not connected, whereas in the latter segments the points are connected. The trajectories are superimposed on a contour plot of the free energy with levels spanning the range from -67.5 to -82.5 in increments of 2.5. (*Right*) A very fast event in which the system moves almost ballistically through the transition region. The last event occurs in roughly 3000 Monte Carlo steps. The trajectories shown were chosen at random from a sample of 86. The sequence used was ABABBBCBACBABABACACBACAACAB and was studied at the folding temperature (T_f = 1.509). The model is a three-dimensional cubic lattice heteropolymer with a contact potential. If the two monomers are of the same type, then the energy for the contact is $E_l = -3$ and if the monomers are not the same the energy is $E_u = -1$. The above sequence was designed to have an unfrustrated nondegenerate native states; i.e., in the native state all contacts are between monomers of the *same* type.

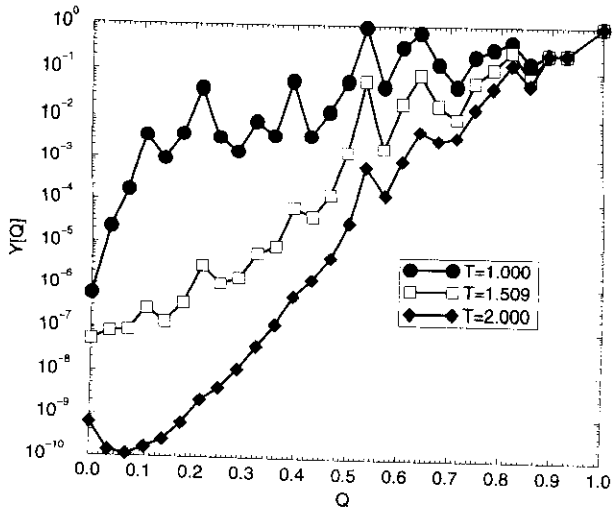


FIG. 4. A plot of $Y(Q) = \Sigma_i[P_i(Q)]^2$ vs. Q for three temperatures. While at the global T_g discrete states are apparent even for small degrees of nativeness, at $T_f = 1.509$ the discrete intermediate is highly native-like.

to the local glass transition may explain the mutation sensitivity of some collapsed globules (17).

Discussion

Experimental data along with simple geometrically based statistical mechanics help locate small helical proteins in their phase diagram, allowing an estimate of the parameters needed to describe folding by statistical energy landscape analysis. A law of corresponding states relates simple lattice models to the laboratory situation, leading to an outline of the topography of a realistic folding funnel, which can serve as a starting point for other investigations. On the experimental side, our analysis pinpoints a great need for more dynamic measurements on the molten globule state itself, one of the weaker points in the numerical estimates. Also the order of collapse and secondary structure formation still needs resolution. The quantitative features of the funnel should help guide and can be refined by fast folding experiments made possible by laser-induced initiation of folding (45). For the folding funnel of the three-letter code model, the minimal frustration of the protein results from harmony between tertiary contacts. Direct local biases like those in the framework picture (46) can be included as an additional slope to the funnel through A rather than Q. Similarly secondary structure may be more directly coupled to the landscape if effective pair interactions depend specifically on the helicity of segments. These considerations require a still more multidimensional view of the funnel, but the low dimensional picture here can serve as a zeroth order starting point.

The vigorous debates among members of the minimalist folding community including J. Bryngelson, H. S. Chan, K. Dill, E. Shakhnovich, and D. Thirumalai helped us crystallize our thoughts. We also thank W. Eaton and H. Frauenfelder for reading the manuscript. N.D.S. is a University of California at San Diego Chancellor Fellow. This work was supported by the National Institutes of Health (Grant 1R01 GM44557), the Beckman Foundation, and the National Science Foundation (Grant MCB-93-16186).

1. Bryngelson, J. D. & Wolynes, P. G. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 7524–7528.
2. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 4918–4922.
3. Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 9029–9033.
4. Leopold, P. E., Montal, M. & Onuchic, J. N. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 8721–8725.
5. Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G. (1994) *Proteins*, in press.
6. Levitt, M. & Warshel, A. (1975) *Nature (London)* **253,** 694–698.
7. Honeycutt, J. & Thirumalai, D. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 3526–3529.
8. Abe, H. & Gō, N. (1980) *Biopolymers* **20,** 1013–1031.
9. Skolnick, J. & Kolinski, A. (1991) *J. Mol. Biol.* **221,** 499–531.
10. Chan, H. S. & Dill, K. A. (1991) *Annu. Rev. Biophys. Biophys. Chem.* **20,** 447.
11. Shakhnovich, E., Farztdinov, G., Gutin, A. M. & Karplus, M. (1991) *Phys. Rev. Lett.* **67,** 1665–1668.
12. Miller, R., Danko, C. A., Fasolka, M. J., Balazs, A. C., Chan, H. S. & Dill, K. A. (1992) *J. Chem. Phys.* **96,** 768–780.
13. Lifshitz, E. M. & Pitaevskii, L. P. (1980) *Statistical Physics* (Pergamon, Oxford), 3rd Ed.
14. Wilson, K. G. & Kogut, J. (1974) *Phys. Rep.* **12,** 75–200.
15. Ptitsyn, O. B. (1992) in *Protein Folding*, ed. Creighton, T. E. (Freeman, New York), p. 243.
16. Griko, Y. V., Privalov, P. L., Venyaminov, S. Y. & Kutyshenko, V. P. (1988) *J. Mol. Biol.* **202,** 127–138.
17. Lin, L., Pinker, R. J., Forde, K., Rose, G. D. & Kallenbach, N. R. (1994) *Nat. Struct. Biol.* **1,** 447–452.
18. Flanagan, J. M., Kataoka, M., Fujisawa, T. & Engelman, D. M. (1993) *Biochemistry* **32,** 10359–10370.
19. Luthey-Schulten, Z. A., Ramirez, B. E. & Wolynes, P. G. (1995) *J. Phys. Chem.* **99,** 2177–2185.
20. Bascle, J., Garel, T. & Orland, H. (1993) *J. Phys. (Paris)* **3,** 245–253.
21. Bashford, D., Karplus, M. & Weaver, D. L. (1990) in *Protein Folding*, eds. Gierasch, L. M. & King, J. (Am. Assoc. Advance. Sci., Washington, DC), pp. 283–290.
22. Kim, P. S. & Baldwin, R. L. (1990) *Annu. Rev. Biochem.* **59,** 631–660.
23. Flory, P. J. (1969) *Statistical Mechanics of Chain Molecules* (Wiley, New York).
24. McCammon, J. A. & Harvey, S. C. (1987) *Dynamics of Proteins and Nucleic Acids* (Cambridge Univ. Press, New York).
25. Doi, M. & Edwards, S. F. (1986) *The Theory of Polymer Dynamics* (Oxford Univ. Press, Oxford).
26. Bryngelson, J. D. & Wolynes, P. G. (1989) *J. Phys. Chem.* **93,** 6902–6915.
27. Baum, J., Dobson, C. M., Evans, P. A. & Hanly, C. (1989) *Biochemistry* **28,** 7–13.
28. Feng, Y., Sligar, S. G. & Wand, A. J. (1994) *Nat. Struct. Biol.* **1,** 30–36.
29. Bryngelson, J. D. & Wolynes, P. G. (1990) *Biopolymers* **30,** 177–188.
30. Frauenfelder, H., Alberding, N. A., Ansari, A., Braunstein, D., Cowen, B., Hong, M., Iben, I., Johnson, J., Luck, S., Marden, M. Mourant, J., Ormos, P., Reinisch, L., Scholl, R., Shyamsunder, E., Sorensen, L., Steinbach, P., Xie, A.-H., Young, R. & Yue, K. (1990) *J. Phys. Chem.* **94,** 1024–1037.
31. Frauenfelder, H. & Wolynes, P. G. (1994) *Phys. Today* **47,** 58–64.
32. Socci, N. D. & Onuchic, J. N. (1994) *J. Chem. Phys.* **101,** 1519–1528.
33. Sfatos, C. D., Gutin, A. M. & Shakhnovich, E. I. (1994) *Phys. Rev. E* **50,** 2898–2905.
34. Yue, K., Fiebig, K., Thomas, P. D., Chan, H. S., Shakhnovich, E. I. & Dill, K. A. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 325–329.
35. Sali, A., Shakhnovich, E. & Karplus, M. (1994) *Nature (London)* **369,** 248–251.
36. Thirumalai, D. & Guo, Z. (1995) *Biopolymers* **35,** 137–140.
37. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1994) *Biochemistry* **33,** 10026–10036.
38. Ferrenberg, A. M. & Swendsen, R. H. (1988) *Phys. Rev. Lett.* **61,** 2635–2638.
39. Hansmann, U. H. E. & Okamoto, Y. (1993) *J. Comput. Chem.* **14,** 1333–1338.
40. Hao, M.-H. & Scheraga, H. A. (1994) *J. Phys. Chem.* **98,** 4940–4948.
41. Peng, Z.-Y. & Kim, P. S. (1994) *Biochemistry* **33,** 2136–2141.
42. Otzen, D. E., Itzhaki, L. S., ElMasry, N. F., Jackson, S. E. & Fersht, A. R. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 10422–10425.
43. Mezard, M., Parisi, G. & Virasoro, M. A. (1987) *Spin Glass Theory and Beyond* (World Scientific, Singapore).
44. Shakhnovich, E. & Gutin, A. (1990) *J. Chem. Phys.* **93,** 5967–5971.
45. Jones, C. M., Henry, E. R., Hu, Y., Chan, C.-K., Luck, S. D., Bhuyna, A., Roder, H., Hofrichter, J. & Eaton, W. A. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 11860–11864.
46. Kim, P. S. & Baldwin, R. L. (1982) *Annu. Rev. Biochem.* **51,** 459–489.

PROTEINS: Structure, Function, and Genetics 21:167–195 (1995)

RESEARCH ARTICLES

Funnels, Pathways, and the Energy Landscape of Protein Folding: A Synthesis

Joseph D. Bryngelson,[1] José Nelson Onuchic,[2] Nicholas D. Socci,[2] and Peter G. Wolynes[3]
[1]Physical Sciences Laboratory, Division of Computer Research and Technology, National Institutes of Health, Bethesda, Maryland 20892, [2]Department of Physics-0319, University of California at San Diego, La Jolla, California 92093-0319, and [3]School of Chemical Sciences and Beckman Institute, University of Illinois, Urbana, Illinois 61801

ABSTRACT **The understanding, and even the description of protein folding is impeded by the complexity of the process. Much of this complexity can be described and understood by taking a statistical approach to the energetics of protein conformation, that is, to the energy landscape. The statistical energy landscape approach explains when and why unique behaviors, such as specific folding pathways, occur in some proteins and more generally explains the distinction between folding processes common to all sequences and those peculiar to individual sequences. This approach also gives new, quantitative insights into the interpretation of experiments and simulations of protein folding thermodynamics and kinetics. Specifically, the picture provides simple explanations for folding as a two-state first-order phase transition, for the origin of metastable collapsed unfolded states and for the curved Arrhenius plots observed in both laboratory experiments and discrete lattice simulations. The relation of these quantitative ideas to folding pathways, to uniexponential vs. multiexponential behavior in protein folding experiments and to the effect of mutations on folding is also discussed. The success of energy landscape ideas in protein structure prediction is also described. The use of the energy landscape approach for analyzing data is illustrated with a quantitative analysis of some recent simulations, and a qualitative analysis of experiments on the folding of three proteins. The work unifies several previously proposed ideas concerning the mechanism protein folding and delimits the regions of validity of these ideas under different thermodynamic conditions**

Key words: protein folding, energy landscape, folding pathway, folding funnel, lattice simulation, folding thermodynamics, folding kinetics, protein engineering

INTRODUCTION

The apparent complexity of folded protein structures and the extraordinary diversity of conformational states of unfolded proteins make challenging even the description of protein folding in atomistic terms. Soon after Anfinsen's classic experiments on renaturation of unfolded proteins,[1] Levinthal recognized the conceptual difficulty of a molecule searching at random through the cosmologically large number of unfolded configurations to find the folded structure in a biologically relevant time.[2] To resolve this "paradox," he postulated the notion of a protein folding pathway. The search for such a pathway is often stated as the motive for experimental protein folding studies. On the other hand, the existence of multiple parallel paths to the folded state has been occasionally invoked.[3] Recently, a new approach to thinking about protein folding and about these issues specifically has emerged based on the statistical characterization of the energy landscape of folding proteins.[4–6]

This paper presents the basic ideas of the statistical energy landscape view of protein folding and relates them to the older languages of protein folding pathways. The use of statistics to describe protein physical chemistry is quite natural, even though each protein has a specific sequence, structure, and function essential to its biological activity. The huge number of conformational states immediately both allows and requires a statistical characterization. In addition folding is a general behavior common to a large ensemble of biological molecules. Many different sequences fold to essentially the same structure as witnessed by the extreme dissimilarities in sequence which may be found in families of proteins such as lysozyme.[7] Thus for any given

Received July 7, 1994; revision accepted November 4, 1994.
Address reprint requests to Joseph D. Bryngelson, Physical Sciences Laboratory, Division of Computer Research and Technology, National Institutes of Health, Bethesda, MD 20892.

observed protein tertiary structure, there is a statistical ensemble of biological molecules which fold to it. Many studies suggest that the dynamics of many parts of the folding process are common to all of the sequences of a given overall structure, while others are peculiar to individual sequences. *Distinguishing folding processes common to all sequences from those peculiar to individual sequences is a major goal of physical theories of protein folding.* The statistical energy landscape analysis will show which features are common and which are specific taxonomic aspects of protein folding.

Depending on the statistical characteristics of the energy landscape, either a unique folding pathway or multiple pathways may emerge. A biological relevance of the distinction between the two pictures is that mutations can more dramatically affect the dynamics through unique pathways than through multiple pathways.

The organization of this paper is as follows: in the next section we describe the energy landscape of protein folding, discuss the properties of smooth and rough energy landscapes, and indicate that it appears that protein folding occurs on an energy landscape that is intermediate between most smooth and most rough. In the third section we describe a simple protein folding model that interpolates between these two limits and exhibits both the smooth and the rough energy landscape properties that are present in folding proteins. The equilibrium thermodynamic properties of this model are also discussed in this section. The fourth section starts with a short survey of the differences between the kinetics of complex chemical processes, such as protein folding, and the kinetics of the simple chemical processes whose understanding forms the basis of the most commonly used reaction rate theories. We review how these common theories should be modified to cope with the complexity of a process like protein folding. Then we present the necessary modifications of kinetics and apply them to the simple protein folding model of the third section. Each scenario has its own characteristic behavior. The folding scenario observed in any given experiment depends on the specific sequence and the refolding conditions. The fifth section shows how the scenarios presented in the fourth section can be understood in terms of the phase diagram for protein folding. This phase diagram is also discussed in detail. In the next section we show how the energy landscape ideas can be used to analyze data by presenting a rough but quantitative analysis of some computer simulation data. In the seventh section, we give a flavor of the issues in energy landscape analysis of experimental data through an examination of some previously published experimental results. We also present a tentative assignment of the folding scenarios observed in these experiments. The concluding section then summarizes the results, and discusses the significance of the energy landscape for understanding protein folding, for protein structure prediction, and for protein engineering.

SMOOTHNESS, ROUGHNESS, AND THE TOPOGRAPHY OF ENERGY LANDSCAPES

Protein folding is a complex process, typically occurring at a constant pressure and temperature, involving important changes in the structure of both the chain and the solvent.[8,9] The natural thermodynamic potential for describing processes at constant pressure and temperature is the Gibbs free energy,[10,11] so we will use an effective free energy that is a function of the configuration of the protein to describe the protein–solvent system. Notice that this description implicitly averages over the solvent coordinates. This averaging means that the forces that arise from this potential function are temperature dependent. To make these considerations more concrete, consider the forces on two apolar groups immersed in water. The apolar group–solvent system has a lower free energy if the two apolar groups are close to one another, so the solvent-averaged free energy, mentioned above, has a minimum when the two groups are close and becomes larger when the groups are further apart.[12] The change in the solvent-averaged free energy as a function of distance between the groups causes the groups to attract one another. This attraction is the hydrophobic force. Since the free energy of the apolar group–solvent system changes as the temperature changes, likewise the solvent-averaged free energy and the hydrophobic force also change.[12,13]

The need to consider the form of the free energy as a function of protein conformation, which we call the energy landscape, stems, in part, from a well-known argument of Cyrus Levinthal.[2] The argument starts by noticing that the number of possible conformations in a protein scales exponentially with the number of amino acid residues. Thus, if each amino acid has only two possible conformations, then the number of possible conformations for a protein with 100 amino acids is $2^{100} \approx 10^{30}$. If, as a conservative estimate, at least 1 ps is required to explore each conformation, then the time required to explore all conformations of the 100 amino acid protein is approximately 10^{18} s, or more than 10^{10} years. From this estimate Levinthal argued that the protein did not have enough time to find its global free energy minimum, so the final, folded conformation of a protein must be determined by kinetic pathways. This argument is easily criticized. For example, one could equally well apply it to the formation of crystals, and conclude that crystallization can never occur! More seriously, the argument can be used to question how the protein could reliably find *any* particular conformation. In this form the argument is often called Levinthal's paradox. The weak point in Levinthal's argument is the assumption that all con-

formations are equally likely in the path from the unfolded to the folded states. In fact, conformations with lower free energy are more likely than those with higher free energy. Levinthal's argument assumes a free energy landscape that looks like a flat golf course with a single hole at the free energy minimum. The argument breaks down completely for a free energy landscape that looks like a funnel.[5,14–16] A central purpose of this paper is to further develop this intuitive notion of energy landscape and to describe quantitatively kinetic behavior on the kinds of energy landscapes that are encountered in protein folding. Interestingly, Levinthal's paradox will reoccur, albeit in a completely different form.

The most detailed description of the energy landscape of a folding protein molecule would be obtained by specifying the free energy averaged over the solvent coordinates as a function of the coordinates of every atom in the protein. At this fine level of description, the free energy surface of a protein is riddled with many local minima.[17,18] Most of these minima correspond to small excitation energies connected with individual local conformational changes such as rotations of individual side chains. The energies involved in these small conformational changes are typically on the order of k_BT, that is, the size of the thermal energies of the atoms in the protein. Interconversion between these shallow local minima will be rapid on the time scale of protein motions. Sometimes many side chains can shift, giving quite different minima with a large energy barrier between them. Changes of backbone conformation can lead to globally different protein folds involving many different interresidue contacts. The energies involved in these larger conformational differences can easily become many times k_BT, and interconversion between these deeper, globally different local minima can be quite slow.[17,18]

The interesting features of protein folding dynamics concern the free energy surface viewed on this more coarse-grained structural scale. Very different behavior occurs, depending on whether this coarse-grained energy landscape is "smooth" or "rough." In Figure 1 we show representative smooth and rough energy landscapes. A smooth energy landscape has only a small number of deep valleys and/or high hills. For smoother energy landscapes there are typically many high energy structures and only a few low energy structures. The more closely the system resembles a few low energy structures, the lower the energy. Thus, each of the low energy structures is at the bottom of a broad energy valley. A protein molecule that was in one of the valleys would find itself dynamically funneled to the lowest energy state. Therefore, we will refer to the valley associated with a low energy structure as a "funnel." In this language, a system with a smooth energy landscape has a few deep minima, each having a large, broad funnel. Systems with smooth landscapes exhibit cooperative phase transitions, illustrated by such phenomena as crystallization of simple materials and in biological macromolecules by phenomena such as the helix–coil transition.[19] The thermodynamic phases of systems with smooth energy landscapes are determined by the temperature. At high temperatures, the large number of high energy structures predominate, but as the temperature of the system is lowered, the system will occupy the lower energy states. Dynamically, below a transition temperature, such systems will fall into a funnel of low energy states and may remain trapped there. In typical cooperative transitions such as crystallization, once a large enough nucleus of low energy structure is formed, the rest of the low energy structure forms rapidly.[20–22]

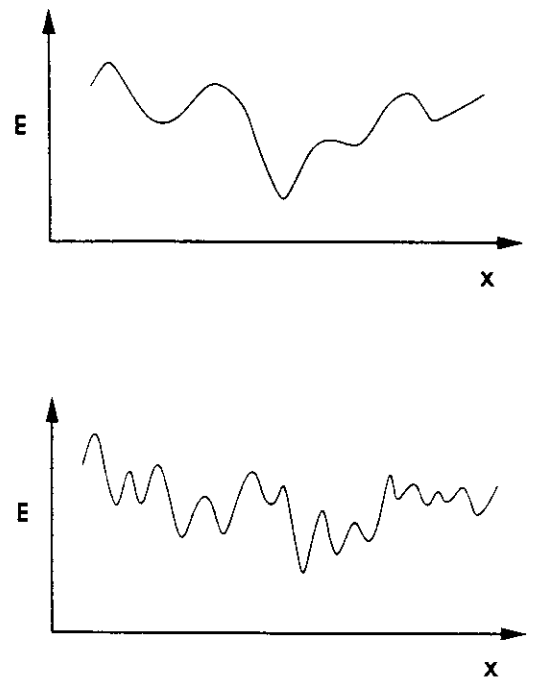


Fig. 1. The energy of a system (vertical axis) is sketched against a single coordinate (horizontal axis) for representative smooth and rough energy landscapes. The top sketch shows a smooth landscape with only a few energy minima each having a broad funnel leading to it. The bottom sketch shows a rough energy landscape with many energy minima each with a narrow funnel leading to it.

Thermodynamically, protein tertiary structure formation for smaller proteins has been shown to exhibit this type of cooperative behavior. For small, single domain proteins, at most two states are observed on the longest time scales under physiological solvent conditions: One a high entropy high energy disordered phase corresponding to the unfolded protein, and a lower entropy low energy phase describing the folded protein.[23] The fact that the phase

space is divided into two main parts is confirmed by the coincidence of transitions measured by different probes such as optical rotation or fluorescence.[24–26] In addition on the longest time scales, one sees only a single exponential in the kinetics of folding. Simulations of protein folding have shown evidence of nucleation-like behavior.[27] Thus these aspects of tertiary structure formation are characteristic of a system with a smooth energy landscape.

Smooth energy landscapes are so commonly used in the description of problems that systems with rough energy landscapes are considered exotic and have only recently been studied by chemists and physicists. A rough energy landscape would be one that, when coarse grained, has many deep valleys and very high barriers between them. In such a rough energy landscape there are a very large number of low energy structures that are entirely different globally. Each of these diverse low energy structures has a small funnel leading to it.

The thermodynamic and kinetic behavior of systems having rough energy landscapes is quite distinct from those with smooth landscapes. Rough energy landscapes occur in problems in which there are many competing interactions in the energy function. This competition is called "frustration." The paradigm for a frustrated system is the spin glass, a magnetic system in which spins are randomly arrayed in a dilute alloy.[28–30] The interactions between spins are equally often, and at random, ferromagnetic (the spins want to point in the same direction) and antiferromagnetic (the spins want to point in opposite directions). These two conflicting local tendencies (one to parallel spins, the other to alternating spins) cannot be satisfied completely in any arrangement of spin orientations. Thus, the system is said to be "frustrated."[31] Many optimization problems that arise in economic contexts have rough energy landscapes because of frustrated interactions. An economic example of a rough landscape is provided by the traveling salesman problem. In this problem one attempts to minimize the total length of a journey which visits each of a set of randomly arrayed cities precisely once during the trip. Here searching for the minimum length trip leads to an optimization problem in which there are many alternate routes that have very nearly the same value of the required length (equivalent to multiple minima). The frustration here arises from the constraint of a single visit to a city because of an occupancy tax; no central location can be used as a base. Finding the optimal solution to this problem is a difficult task. Computer scientists have developed a set of ideas that describes many problems that are hard to solve.[32] Although the precise technical framework of these ideas is elaborate, the basic idea is simple; there exists a set of difficult problems that cannot be solved by any known polynomial time algorithm, and it is generally believed that no such algorithm exists. These problems are called NP-complete. Here by polynomial time algorithm we mean that the amount of computation time required to solve the problem grows no faster than some fixed power of the problem size, e.g., the number of cities in the traveling salesman problem. Furthermore, the general model of computation used in NP-completeness proofs is thought to be able to simulate any natural system, so the limitations that NP-completeness impose on computation probably hold for all natural systems, e.g., folding proteins, the human brain etc.. Thus, solutions to NP-complete problems require an exponential, rather than polynomial, amount of time. In practical terms, NP-completeness means that the amount of time required to solve even modest size problems can become astronomically large. The traveling salesman problem is an example of an NP-complete problem; that is, its solution for the general case requires exponentially more computational time as the size of the problem grows. Finding the lowest free energy state of a macromolecule with a general sequence also has been shown to be NP-complete.[33] NP-completeness is a worst case analysis; if a problem is proven to be NP-complete then finding the solution to at least one case requires an exponential computation time. In economic situations these computational difficulties are avoided by choosing to be satisfied with an acceptable solution *or* by selecting the conditions of the problem so that easy answers can be found. An example of the latter is the introduction of the "hub" system to airline traffic. A central city, perhaps not usually visited, is introduced as a place that can be multiply visited at little cost. Similarly for the physicist's spin glass, there are some specifically chosen arrangements of ferromagnetic and antiferromagnetic interactions so that each interaction can be satisfied in a single configuration. The arrangements of interactions which do this are relatively improbable. Therefore, in the context of proteins, NP-completeness means that there are amino acid sequences that *cannot* be folded to their global free energy minimum in a reasonable time either by computer or by the special algorithm used by nature. Thus, in analogy with the economic situation, either naturally occurring proteins fold to a structure that is not a global minimum *or* they have been selected to be members of the subset of amino acid sequences that *can* fold to their global free energy minimum in a reasonable time. The NP-completeness proof alone does not distinguish between these two possibilities. If the latter possibility is correct then one approach to predicting structure is simulated annealing.[34] Starting at high temperature, the system is slowly brought to low temperature while following its dynamics. These stochastic search algorithms parallel the Levinthal paradox for protein folding kinetics. Such an approach can work only if the computer's energy landscape is sufficiently close to the one that nature used.

In any case, even if proteins fold to a structure that is not a global minimum, i.e., folding is kinetically controlled, they must *reliably* fold to a single structure. Recent experiments on random and designed amino acid sequences have shown that reliable folding is not a universal property of polypeptide chains, and that multiple folded structures are the rule rather than the exception.[35,36] Thus both theory and experimental evidence indicate that such reliable folding characterizes only a small fraction of amino acid sequences. Proteins are a subset of this fraction of reliable folders. Later in this paper we discuss a property we call minimal frustration. Evidence from theory and from simulation indicates that amino acid sequences with minimal frustration are likely to fold reliably.

What are the possible sources of frustration in the general case of a heteropolymer? Consider the hydrophobic effect, which for illustration we think of as a contact interaction favoring hydrophobic pairs or hydrophilic pairs. Because of the constraint of chain connectivity for most random sequences bringing together a hydrophobic pair distant in sequence will require bringing together other pairs in the sequence which will often be dissimilar and therefore unfavorable.* This situation could be avoided in natural proteins by choosing simple patterns of hydrophobic–hydrophilic alternation like those seen in β-barrel proteins.[37] Similarly, for most sequences the hydrophobicity pattern favoring a particular secondary structure (α-helix or β-sheet) might or might not be consistent with the tendency of each amino acid to be in that secondary structure. Indeed, in general there is usually some conflict of this sort, since the ends of α-helices have unsatisfied hydrogen bonds, but the helices must be broken so that a compact structure can form, satisfying the hydrophobic forces. Sequences may need special start or stop residues to form terminal hydrogen bonds gracefully, using side chains.[38,39]

Polymers can also exhibit another kind of frustration. A molecule often needs to overcome an energy barrier to change from one structure to another. This notion has been used explicitly in the simulation studies of Camacho and Thirumalai and of Chan and Dill where they constructed paths with minimal energy barriers between similar configurations in their protein folding models and used this network of pathways to map out several features of the energy landscape.[40–42] If this energy is too high to overcome in a reasonable time, for example, some fraction of the folding time for a protein, then we may say that the two structures are not "dynamically connected." Two different structures may resemble each other, and even have similar free energies, but they may be unable to reconfigure from one to the other one in any reasonable time scale. Such structures would not be dynamically connected. In particular, for polymers, geometric constraints arise because the polymer chain cannot pass through itself. This effect is called excluded volume, and may give rise to an enormous energy barrier. In this case one can easily have two structures that resemble each other but are not dynamically connected. Leopold et al. have explicitly shown that this situation occurs in some simple models of protein folding.[15] We will refer to this kinetic phenomenon as geometric frustration.

Systems with rough energy landscapes also exhibit effective phase transitions.[28–30] When the temperature of such a system is lowered, it tends to occupy the lower energy states and at a transition temperature will become trapped in one of them. Generally, these transitions are accompanied by a considerable slowing of the motion as the system tries to exit over the high energy barriers. In the case of liquids being supercooled below their freezing point, this phenomenon is known as the glass transition.[43,44] Below the glass transition temperature, the liquid is trapped in a single deep minimum and thus it looks like a solid. The thermodynamics of this solid depends on its detailed thermal history. Typically, systems with rough energy landscapes exhibit glass transitions analogous to those that occur when liquids are supercooled below their freezing point. As the system approaches the glass transition, the slow transitions between minima leads to strongly nonexponential time dependences for many properties.

Typically a heteropolymer with a random sequence interacting with itself has a rough energy landscape. One source of the roughness is the frustration arising from conflicting interactions but geometric constraints may be important too. In either case, energetic or geometric frustration, there will be a large barrier to reconfigure between these configurations. This is a natural starting assumption for thinking about heteropolymer dynamics since one expects this behavior generically for heterogeneous systems. The implications of the roughness of heteropolymer energy landscapes for protein folding were first discussed by Bryngelson and Wolynes who postulated that the energies of the states of a random heteropolymer could be approximately modeled by a set of random, independent energies.[4] This model is known as the random energy model in the theory of spin glasses.[45–47] The random energy model approximation used by Bryngelson and Wolynes was later shown to be equivalent to a more conventional replica mean field approximation by Garel and Orland[48] and by Shakhnovich and Gutin.[49] A direct demonstration of the roughness of the energy landscape for heteropolymers has been carried out for small lattice model proteins with simple

*It is useful for the reader to study Figure 2, in which we illustrate the varying degrees of frustration for two sequences of a lattice model of a heteropolymer.

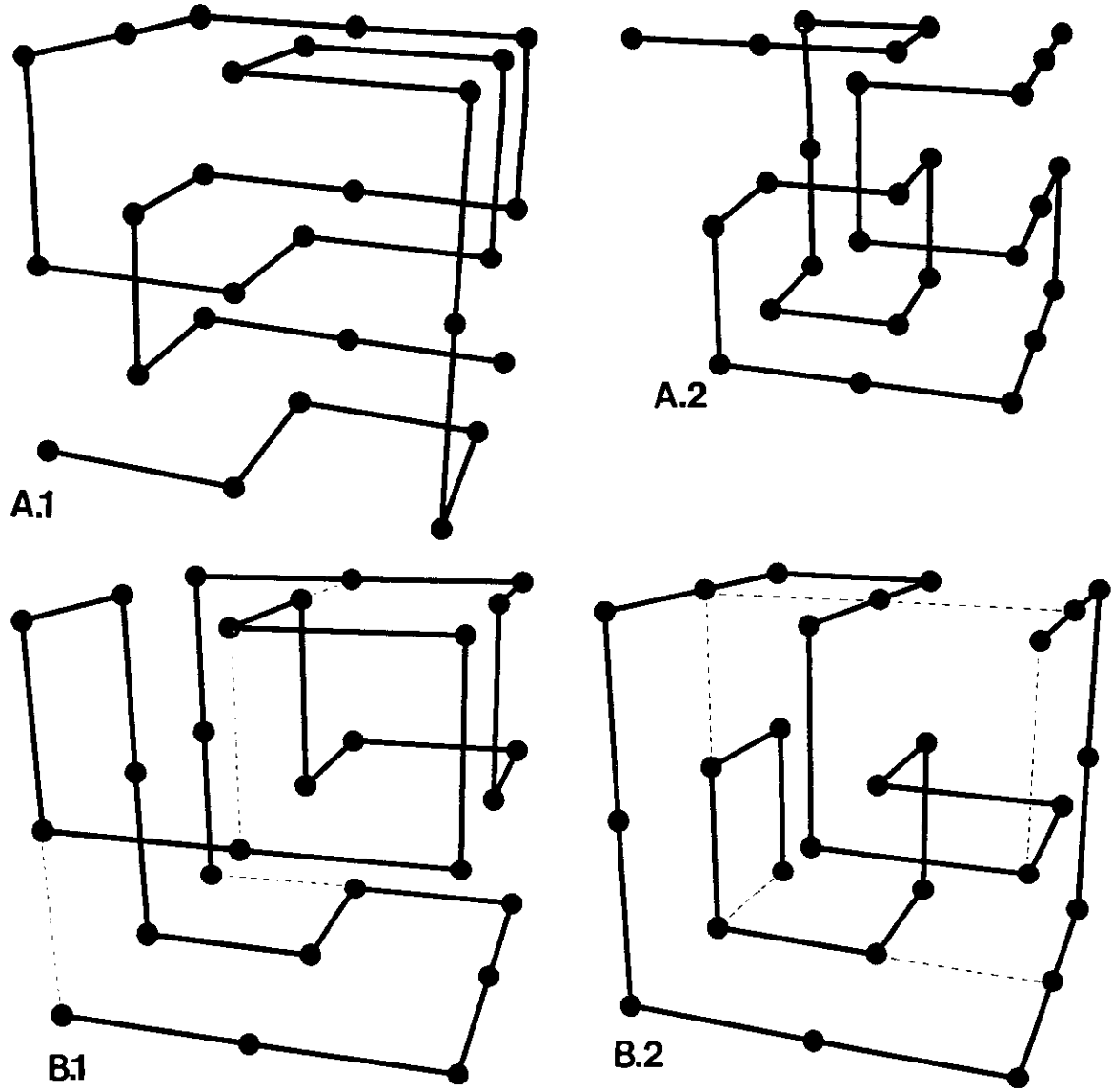


Fig. 2. The ground state of two different sequences for a 27-mer, with two different types of monomers (two letter code) on a cubic lattice. If two monomers are adjacent in space, but not along the chain, then there is an attractive interaction between them. This interaction is strong if the monomers are of the same type and weak if they are of different types. For all figures we use the following notation: solid lines represent covalent bonds, dashed lines represent spatial contacts with weak interactions, and no lines are drawn for spatial contacts with strong interactions. The model for this 27-mer is presented fully in the section on Energy Landscape Analysis of Folding Simulations. The strong interactions are equal to -3 and the weak ones to -1 in arbitrary energy units. The most compact configurations will be cube-like and they have 28 spatial (nonbonded) contacts. Sequence (A) has only strong contacts in its ground state. For this reason we call it a nonfrustrated ground state. **(A1)** The ground state structure for this sequence. We call it nonfrustrated because all contacts are optimal. We show in the section on Folding Simulations that this sequence is a good folding sequence. This is not the case for sequence (B). Its ground state configuration has 4 weak interactions, as shown in **B1**. For this reason we say that this sequence is frustrated, i.e., it is unable to optimize all the interactions and it has to compromise with some weak ones. We show in the section on Folding Simulations that sequence (B) is not a good folder. However, there is a more interesting way of observing frustration. Let us call Q a measure of similarity between the ground state configuration and any other configuration (compact or noncompact) for a given sequence. The quantity Q measures the number of contacts between pairs of residues that are the same for a given configuration and its ground state one. Therefore, Q is a number between 0 and 28. Most of the configurations with energy just above the ground state in sequence (A) have Q between 18 and 26, i.e., very similar to the ground state configuration. An example of such a configuration is shown in **A2** where the energy is -78 and Q is 26. The situation is completely different for sequence (B). There are configurations with energy just above the ground state configuration that have a Q between 4 and 12, i.e., they are very different from the ground state. An example of one of these configurations is shown in **B2** where the energy is -72 and Q is 9. In this case, there are lots of low energy states that are completely different but energetically very similar. When the system gets trapped in one of these low energy states, it takes a long time to completely reconfigure before it can try to fold again.

interactions. Here the exact enumeration of configurations can be carried out and it can be directly established that there are configurations very close in energy to the ground state that have topologically distinct folds for most random sequences.[50-55] Work on realistic lattice models for small proteins confined to their proper shape (where complete enumeration can be carried out) suggests the possibility of deep low energy structures that are globally different in form.[56,57] Even for a well designed sequence (i.e., one designed to have a smooth energy landscape) some roughness may remain. Early direct evidence for roughness in the energy landscape of protein folding simulations of designed sequences is provided by the work of Honeycutt and Thirumalai, who looked for and found deep multiple energy minima in their simulations of β-barrel folding.[58,59] Finally, the historical difficulty of predicting protein structure from sequence arises from the "multiple minimum problem," that is, the existence of many minima in the empirical potential energy functions used to predict these structures. The large number of minima indicates that the energy landscape of these potential functions is rough. The importance of the multiple minimum problem, and therefore the roughness of the energy landscape, as an impediment to structure prediction has been emphasized by Scheraga and collaborators.[60]

Some experimental features of protein folding suggest a considerable roughness to the energy landscape. Although protein folding appears to be exponential in time, short time scale measurements show the existence of intermediates. Also, multiexponential decay of relaxation properties is seen in these early events.[61] Many of the time scales involved in protein unfolding have very large apparent activation energies, suggesting high energy barriers. There is the occasional report of history dependence to protein folding, although, this is absent from studies on smaller proteins in vitro.[23]

QUANTITATIVE ASPECTS OF THE STATISTICS AND THERMODYNAMICS OF A FOLDING PROTEIN

In the previous section we found that a folding protein exhibits behaviors that are characteristic of both smooth and rough energy landscapes. Thus, from the phenomenological viewpoint it is evident that protein folding occurs on an energy landscape that is intermediate between the most smooth and the most rough. A simple model proposed by Bryngelson and Wolynes interpolates between the two limits and illustrates the basic ideas of the energy landscape analysis of protein folding.[†4] When stripped down to its bare essentials, this picture of the folding landscape is based on two postulates: The first captures the rough aspects of the energy landscape. It is postulated that (for natural proteins) the energy of a contact between two residues which does not occur in the final native structure of a protein or the energy of a residue in a secondary structure which does not turn out to be ultimately correct can be taken as random variables; that is, in its nonnative interactions, a protein resembles a random heteropolymer. In its extreme form this suggests that we can take the energies of globally distinct states to be random variables which are uncorrelated, provided no native contacts are made and no native secondary structure is formed. A second postulate captures the smooth aspects of the folding landscape. When a part of the protein molecule is in its correct secondary structure, the energy contributions are expected to be stabilizing. In addition, when a correct contact is made, although occasionally the energy may go up, on the average over all possible contacts, the energy will go down. Thus if the similarity to the native structure is used as a distance measure, the surface may have bumps and wiggles but the energy generally rises as we move away from the native structure. Thus there is an overall energetic funnel (of the sort discussed in the previous section) to the native structure.

Bryngelson and Wolynes used the term "the principle of minimal frustration" in describing the smoothness postulate, insofar as it is what distinguished natural proteins from random heteropolymers. The smoothness of folding landscapes arises from the selection of protein sequences by evolution. If the necessity to maximize the ability of folding quickly is the dominant selection pressure, the smooth part of the energy landscape will be paramount. On the other hand, there are other selection pressures as well. Thus evolution may not be able to remove some frustrated interaction from natural proteins. Indeed, neutral evolution would suggest that randomness and frustration would continue to exist to an extent that allows only adequate stability and kinetic foldability. The minimal frustration of natural proteins is evident in several ways. Examination of X-ray structures shows that side chains are in fact chosen by evolution to make coherent contributions to supersecondary structures. Clear examples are leucine zippers[62] and the β-barrel amphiphilicity mentioned earlier. Symmetric sequences like these often lead to low frustration in symmetric structures. Consistency between secondary structures and global tertiary structures is also important. This is the "principle of structural consistency" enunciated by Gō.[63]

Purely kinetic effects also limit the folding of pro-

[†]In this section we use the word "energy" to describe the free energy of a given complete configuration of the protein. Such a configuration has many solvent configurations consistent with it. Thus our energy landscape has a temperature dependence due to hydrophobic forces. We do not consider this effect when we discuss the pure effects of temperature in this section.

teins. For example, if the minimum energy structure is not dynamically connected (in the sense described in the previous section) to any other low energy structures, then it would be kinetically inaccessible in spite of its low energy. The importance of kinetic effects for protein folding was investigated in the previously mentioned study of Leopold et al.[15] They simulated the folding of two "sequences" with their simple model, one of which folded rapidly to its global energy minimum, the other of which failed to find its global energy minimum in several long runs. Analysis of the dynamic connectivities produced by the two "sequences" showed that the minimum energy structure of the rapidly folding sequence had a rich network of dynamic connections to most of the other low energy structures. In contrast, the minimum energy structure of the other sequence was sparsely connected to other low energy structures.

A figure encompassing the qualitative considerations about the folding landscape is pictured in Figure 3A. Of course, no low dimensional figure can do justice to the high dimensionality of the configuration space of a protein, but one sees that the dominant smooth features of the landscape depend on how close a protein configuration is to the native one and this coordinate is specifically shown as the radial coordinate in the figure. There are a variety of ways of measuring the similarity of a protein structure to the native structure. One can take the fraction of the amino acids residues which are in the correct local configuration. This is a choice used in the original papers of Bryngelson and Wolynes.[4-6] Another possibility for measuring tertiary structure is the fraction of pairs of amino acids which are correctly situated to some accuracy. This measure is related to the distance plots used by crystallographers.[64,65] The similarity measure may also be thought of as a measure of the distance between the two structures, so that similar structures are considered to be close to one another. We denote the similarity of a protein structure to the native structure by n. We will take $n = 1$ to denote complete similarity to the native state and $n = 0$ to denote no similarity to the native state. The radial coordinate in Figure 3A should be thought of as this similarity measure n. The average energy of a state with a certain similarity to the native structure has a value that gets lower as the native structure is approached—thus the overall slope of the energetic funnel. On the other hand the rugged part of the energy landscape means that no individual state has precisely this energy and we can characterize the fluctuations in energy with a given similarity to the native structure by the variance, $\Delta E^2(n)$. The ruggedness of the energy landscape as measured by this variance clearly depends on the compactness of the protein molecule since it arises from improper three-dimensional contacts. In general, the variance may also conceivably decrease as the native structure is approached, but this is not essential for our picture.

The energy of a given state arises from the contributions of many terms, so it is natural to assume that the probability distribution of energies for any similarity to the native structure is given by a Gaussian distribution,

$$P(E) = \frac{1}{\sqrt{2\pi\Delta E^2(n)}} \exp - \frac{[E - \bar{E}(n)]^2}{2\Delta E^2(n)}. \quad (1)$$

The other important feature of the statistical landscape description is the number of conformational states of a protein as we move away from the native structure. The total number of conformational states grows exponentially with the length of the protein. If there are γ configurations per residue, this total number of configurations is $\Omega = \gamma^N$. γ depends on the level of description of the model. It is of order 3, 4, or 5 for the backbone coordinates, but might rise to roughly 10 if the side chain configurations are also included in the analysis. As noted above, the ruggedness of the energy landscape is most important when the protein is compact. The number of compact configurations is considerably smaller than the total number and can be estimated from Flory's theory of excluded volume in polymers.[66,67] $\Omega(R)$ decreases quite considerably as the radius of gyration of the protein falls. For maximally compact configurations of the backbone, $\Omega(R) = \gamma^{*N}$ where γ^* is of the order 1.5.

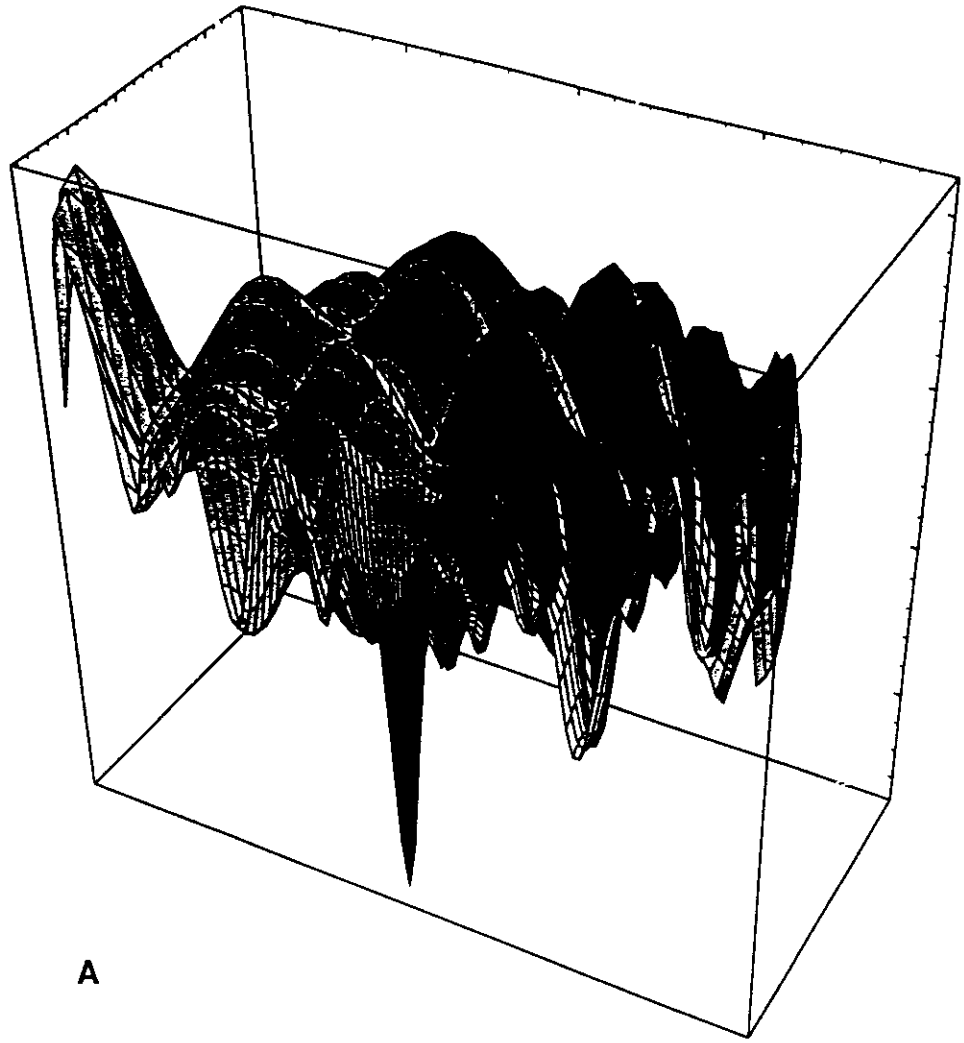
The completely folded protein has a much smaller degree of conformational freedom. Essentially a single backbone structure exists. Thus the number of configurations of the protein decreases as we move toward the native structure. Therefore, if $\Omega(n)$ denotes the number of structures with a similarity measure with the native structure of n, then $\Omega(n)$ and

$$S_0(n) = k_B \log \Omega(n) \quad (2)$$

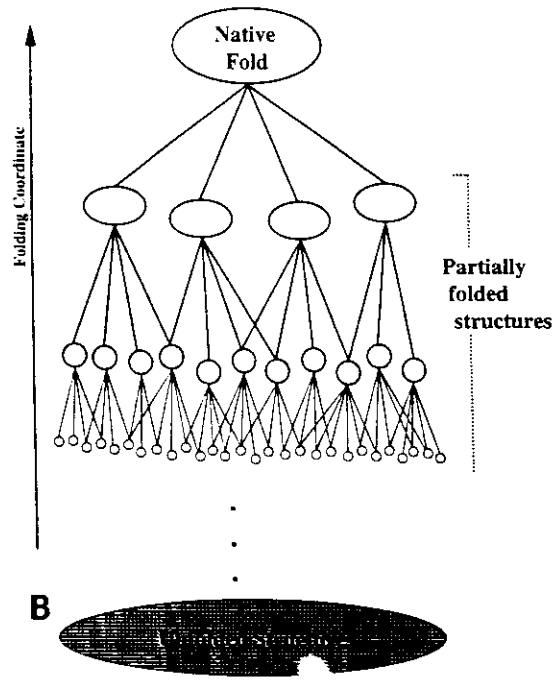
decreases as n gets larger. The exact similarity measure determines the behavior of $S_0(n)$. For our purposes here we need only take a simple form of $S_0(n)$ that decreases as the native state is approached. Roughly speaking, we can approximate Ω as a function of n by $\Omega = \gamma^{*N(1-n)}$.

As one moves away from the native structure there is a huge increase in the number of accessible states, which we can think of as living on the branches of a highly arborized tree as is shown schematically in Figure 3B. Not all of the states on a statistical landscape are thermodynamically or kinetically important, since the high energy states cannot be thermally occupied. The number of states with a specified energy E, which have a specified similarity, n, to the native structure, is given by

$$\Omega(E,n) = \gamma^{*N(1-n)} P(E). \quad (3)$$



A



B

Fig. 3. **(A)** Sketch of an energy landscape encompassing the qualitative considerations about the folding. The energy is on the vertical axis and the other axes represent conformation. This landscape has both smooth and rough aspects. Overall, there is a broad, smooth funnel leading to the native state, but there is also some roughness superimposed on this funnel. Of course, no low dimensional figure can do justice to the high dimensionality of the configuration space of a protein. **(B)** A schematic drawing of protein conformations in relation to their similarity to the native state. The vertical direction is a folding reaction coordinate. The conformations that are higher in the figure are more similar to the native state. As one moves away from the native structure there is a huge increase in the number of possible conformations.

At thermal equilibrium, only a small band of energy is occupied with a certain similarity to the native structure. For a large protein, this band will be relatively well defined in energy. The most probable value of the energy in this band can be obtained by maximizing the thermodynamic weight of states of a given energy. This is a product of the Boltzmann factor and the number of states of that energy

$$p(E,n) = \frac{1}{Z}\Omega(E,n)e^{-E/k_BT}. \quad (4)$$

[Note: Do not confuse $p(E)$ above with the $P(E)$ defined in Eq. (1).] Here Z is the partition function, which ensures normalization of the probability function. Thus the most probable energy with a certain similarity to the native structure is given by

$$E_{m.p.}(n) = \bar{E}(n) - \frac{\Delta E(n)^2}{k_BT} \quad (5)$$

and the number of thermally occupied states is

$$\Omega[E_{m.p.}(n),n] = \exp\left[\frac{S_0(n)}{k_B} - \frac{\Delta E(n)^2}{2(k_BT)^2}\right]. \quad (6)$$

The entropy of the thermally occupied structures that have a certain similarity to the native structure is

$$S[E_{m.p.}(n),n] = k_B \log[\Omega(E_{m.p.}(n),n)]. \quad (7)$$

We see from these expressions that there are two opposing thermodynamic forces involved in the folding process. The growth in the number of thermally occupied states as we move away from the native structure favors a large number of highly disordered configurations. On the other hand, the decreasing average energy as one approaches the native structure favors folded configurations. These two features are combined by thinking of the free energy as a function of the configurational similarity n at a fixed temperature T,

$$\begin{aligned} F(n) &= E_{m.p.}(n) - TS[E_{m.p.}(n),n] \\ &= \bar{E}(n) - \frac{\Delta E(n)^2}{2k_BT} - TS_0(n). \end{aligned} \quad (8)$$

This free energy function is the logarithm of the thermodynamic weight of states with a certain similarity to the native structure. We see in Figure 4 a representation of this free energy and of the probability of occupation. At high temperatures, the band of states with nearly no native structure is favored, corresponding to an unfolded state. At very low temperatures the folded configurations would be favored and, in between, a double minimum effective free energy pertains. The folding temperature is determined by the condition that the two global minima be equal in thermodynamic weight. The unfolded minimum can correspond to two distinct sets of states corresponding to different values of a distinct order parameter, the radius of gyration.[6] If the randomness is large and nonspecific interactions are important, or the chain is highly hydrophobic in composition, this minimum itself can be collapsed. This may well correspond to the molten globule state.[68] On the other hand, if there is little average driving force to collapse due to nonspecific contacts [$\Delta E(n)^2$ small] the disordered configurations will be noncompact and this corresponds to the traditional denatured random coil state. We note that many intermediate degrees of order can exist in the molten globule phase and these can and should be taken into account in a complete analysis. However, the simple one-parameter analysis captures the essentials and should fit data over an appropriately restricted range of thermodynamic conditions.

QUANTITATIVE ASPECTS OF THE KINETICS OF A FOLDING PROTEIN

The theoretical formulation of the kinetics of protein folding differs from the classic formulation of transition state theory in some important ways. Most of our ideas concerning rate theory had their origin in studies of gas phase reactions of small molecules and simple unimolecular reactions in liquids.[69,70] Four important properties of these simple reactions will illustrate the most important points of contrast with protein folding. First, in the simple reactions solvent is either absent or plays a passive role, e.g., as a heat bath or source of friction. Second, the initial state, final state, and transition state all refer to single, fairly well-defined structures so entropy considerations are not important. Third, there is a single, fairly well-defined reaction coordinate. Fourth, the effective diffusion coefficient for moving along the reaction coordinate changes very little as the system moves from the initial to the final state. Protein folding is completely different from these simple reactions.[5,40,41,71] First, in protein folding, the solvent plays a vital role in stabilizing the folded state. As explained above, the important role of the solvent means that the potential of mean force, which here plays the role of the energy as a function of configuration, is a function of temperature and solvent conditions. Second, the initial denatured state, final folded state, and transition states all refer to sets of protein structures, so the configurational entropy of the protein chain is a necessary part of the description of protein folding. Third, there are many possible reaction coordinates and pathways. Fourth, the dynamics of the protein chain changes qualitatively during the course of folding; in particular, an open chain has far greater thermal motion than a collapsed chain. Therefore, the effective diffusion coefficient for motion along a reaction coordinate for folding probably can also change qualitatively between the initial and final states.

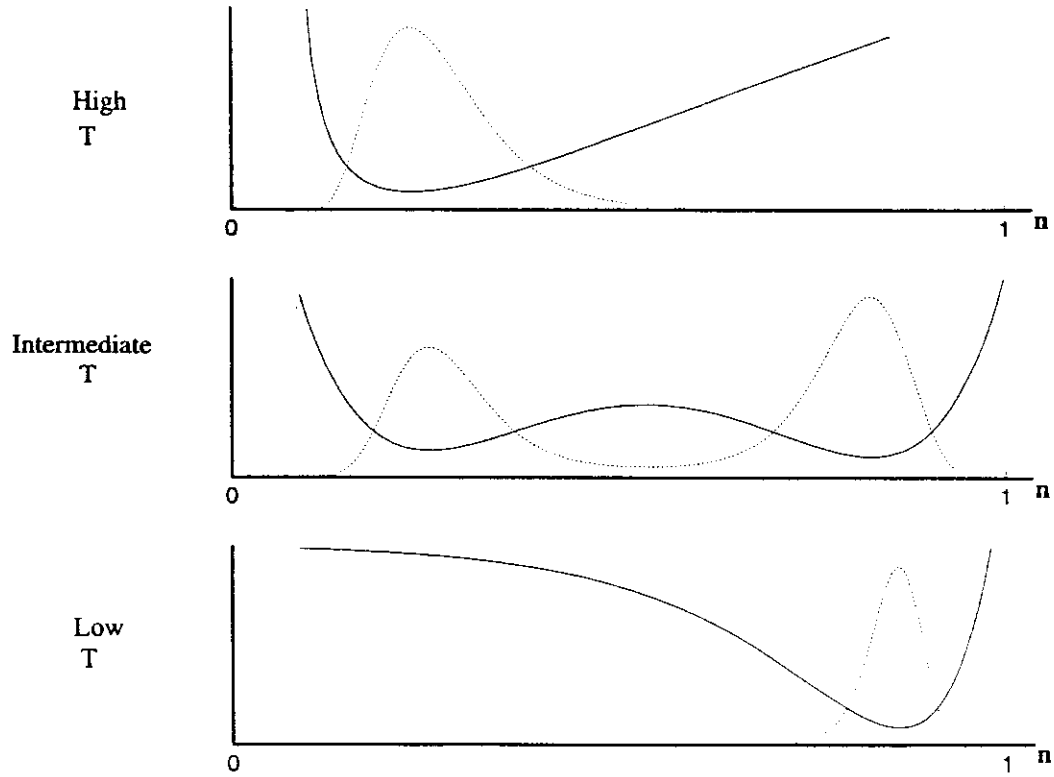


Fig. 4. Sketches of the free energy (solid lines) and probability of occupation (dashed lines) against a folding reaction coordinate for three different temperatures. The value $n = 1$ corresponds to the native structure. The top sketch shows the situation for high temperatures, where the free energy function has a single minimum near $n = 0$, i.e., in highly unfolded states. Here the molecule is far more likely to be in an unfolded conformation than it is to be a conformation similar to the native structure, as it is shown by the dashed lines. As the temperature is lowered, the free energy develops a second minimum, one of them similar to native structure. There is a a free energy barrier between these minima. At these temperatures the probability of occupation is bimodal, with one unfolded and one nearly native peak. Finally at low temperatures, there is again a single minimum in the free energy, but this minimum is near the native structure. Here the molecule is very similar to the native structure.

Below we will discuss the modifications to the transition state theory framework that are needed to describe protein folding kinetics.

The gradient of the free energy function $F(n)$, describes the overall tendency for the system to move and change its value of n. The average flow in configuration space will tend to minimize the free energy. For typical forms of the expressions for the energy and the entropy as a function of similarity to the native state, n, $F(n)$ will tend to have one or two minima, so the system will be unistable or bistable. If the system is unistable and the conditions are favorable for folding, then the single minimum of the free energy function must occur near the native state. A unistable free energy function with its minimum near the native state would require a huge thermal driving force. We call this situation "downhill" protein folding. Downhill folding is rare in slow timescale in vitro protein folding experiments carried out in conditions near the transition between equilibrium folding and equilibrium unfolding. Downhill folding may be common in strongly nativizing conditions, in the initial stages studied in fast timescale folding experiments,[61] and in vivo. In downhill folding the protein folds by making a straight run down the average free energy gradient.

An analogy with transition state theory[69,70] yields a simple estimate for the folding rate, or equivalently, the folding time.[5] In transition state theory the reaction rate is given by the rate of going through the bottleneck for the reaction. Traditionally, this bottleneck is the highest free energy state in the reaction coordinate pathway from the reactant state to the product state. This bottleneck is called the transition state. In transition state theory the rate of going through the transition state depends on the free energy barrier, i.e., the difference between the transition state free energy and the reactant state free energy. In downhill folding there is no free energy barrier. However, there is a bottleneck for folding in downhill folding, because the effective diffusion coefficient for motion along a reac-

tion coordinate changes qualitatively during the course of folding; the region with the smallest diffusion coefficient is the kinetic folding bottleneck. Let $\bar{t}(n)$ denote the typical lifetime of an individual microstate with a similarity n to the native structure. This lifetime is a measure of the rate of motion along the reaction coordinates for folding; the larger $\bar{t}$ the smaller the effective diffusion coefficient and the slower the folding rate. The kinetic bottleneck for folding occurs at the value of n that maximizes $\bar{t}(n)$, which we denote by $n^{\ddagger}_{\text{kin}}$. The subscript kin stands for kinetic and the reason for using this subscript will become apparent below. Therefore, a simple estimate of the folding time, τ, in analogy with transition state theory, is given by

$$\tau = \bar{t}(n^{\ddagger}_{\text{kin}}) . \tag{9}$$

Notice that the time in Eq. (9) is a lower bound on the folding time, hence an upper bound on folding rate. This property is expected because the transition state technique gives upper bounds on reaction rates.[70] We shall discuss the meaning of $n^{\ddagger}_{\text{kin}}$ in more detail below. For now notice that $n^{\ddagger}_{\text{kin}}$ is *not* the location of the top of the free energy barrier, as in conventional transition state theory.

The roughness of the energy surface determines the lifetime of individual microstates. The detailed distribution of these lifetimes can be determined from a detailed analysis and it is rather broad. However, a reasonable first approximation to the typical escape time is easy to obtain. Most minima along a perimeter of constant n are surrounded by ordinary states with nearly the average energy, $\bar{E}(n)$. Thus the barrier height for hopping is $\bar{E} - E_{\text{m.p.}} = (\Delta E/k_{\text{B}}T)^2$. This gives an escape time with a super-Arrhenius temperature dependence

$$\bar{t}(n) = t_0 e^{(\Delta E(n)/k_{\text{B}}T)^2} . \tag{10}$$

The prefactor t_0 is the timescale for a typical motion of a large segment of the chain. It depends on local barriers and on the solvent viscosity, which is itself temperature dependent. Isoviscosity studies of protein folding are therefore quite interesting. The non-Arrhenius temperature dependence exhibited here is sometimes called the Ferry law[72] and it describes the slow dynamics of many glassy systems. As expected, increasing the roughness of the energy landscape greatly slows down folding.

What happens to the escape process as the temperature is lowered? The above estimate assumes many channels for escape exist and an average one can be taken. But as the temperature is lowered it becomes preferable to find an unlikely channel with an improbably low barrier. A subtle analysis[5] shows that, for a given value of n, the escape time goes no lower than a "search" time

$$\bar{t} = t_0 e^{S_0(n)/k_{\text{B}}} . \tag{11}$$

This is the average number of steps taken by the protein to find a state of negligible barrier. This is the Levinthal time for searching states at fixed perimeters, i.e., fixed value of n. For a given n this escape time is reached at a temperature

$$T_{\text{g}}(n) = \left[\frac{\Delta E(n)^2}{2k_{\text{B}}S_0(n)}\right]^{1/2} . \tag{12}$$

The analysis of Bryngelson and Wolynes also shows that for $T > T_{\text{g}}(n)$ the protein has kinetic access to representative section of the perimeter (see Fig. 2) so the behavior of a typical protein molecule can be replaced by the behavior of a statistical ensemble. In this case Eqs. (9) and (10) for the folding time are valid.[‡] For $T < T_{\text{g}}(n)$ the protein has kinetic access to very few structures. These structures are not necessarily representative of the statistical ensemble, so the proteins behavior is dominated by the details of its specific energy landscape. In this case Eqs. (9) and (10) for the folding time must be modified. Technically, the kinetic behavior of the protein molecule becomes non-self-averaging, a term we discuss later in this section.

A system with a fixed n also undergoes a thermodynamic second-order phase transition at $T_{\text{g}}(n)$ in which the protein is effectively frozen into one or few of a small number of low energy states. Using Eq. (6), we see that for $T \leq T_{\text{g}}(n)$, the number of thermally occupied states no longer scales exponentially with the size of the protein.[§] Conversely, as a protein folds at a fixed temperature T, the similarity to the native structure, n, becomes larger. However, the entropy, $S[E_{\text{m.p.}}(n),n]$, decreases as n becomes larger, i.e., there are fewer states available to the protein molecule as it approaches its native structure. A typical protein runs out of entropy at some value n_{g} of n. This vanishing of configurational entropy is precisely the previously noted second-order phase transition, this time taking T rather than n to be constant. The critical values of the temperature and the fraction native structure are related by $T = T_{\text{g}}(n_{\text{g}})$, where T is the temperature at which the folding experiment is carried out. In addition, Bryngelson and Wolynes have shown that the glass transition can occur only for a protein that already has collapsed.[6] Therefore, for any given temperature T, for values of $n \leq n_{\text{g}}(T)$ the kinetic description presented in this section is valid and the folding kinetics are self-averaging, and for $n > n_{\text{g}}(T)$ the protein

[‡]The more subtle analysis by Bryngelson and Wolynes shows that the full time dependence of $\bar{t}$ is slightly more complicated: For $T > 2T_{\text{g}}(n)$, $\bar{t}(n) = t_0\exp[(\Delta E(n)/k_{\text{B}}T)^2]$, as in Eq. (10) above, but for $2T_{\text{g}}(n) > T > T_{\text{g}}(n)$, this equation must be modified to $\bar{t}(n) = t_0\exp[S_0(n) + (1/k_{\text{B}}T_{\text{g}}(n) - 1/k_{\text{B}}T)^2\Delta E(n)^2]$.

[§]Applying Eq. (6) literally would imply a thermally accessible perimeter with less than one state because the entropy analysis neglects finite size corrections.

is in the glassy phase, and its kinetics becomes non-self-averaging.

For bistable systems, there are two minima of free energy with a maximum of free energy between them. In folding conditions the minimum close to the native state has a lower free energy than the minimum corresponding to the unfolded state. The free energy barrier for folding, F_{barrier}, is given by the difference between the free energy of the unfolded minimum, $F(n_{\text{uf}})$ and the free energy of the barrier top, $F(n_{\text{th}}^{\ddagger})$. The subscript th stands for thermodynamic and the reason for using it will become clear momentarily. Systems to the right of the top of the free energy barrier, i.e., with $n > n_{\text{th}}^{\ddagger}$, tend to become folded; those to the left, i.e., with $n < n_{\text{th}}^{\ddagger}$, would become unfolded on the average. A straightforward generalization of transition state theory[5] indicates that the overall folding time is given by

$$\tau = \bar{t}(n_{\text{kin}}^{\ddagger})e^{F_{\text{kin}}^{\ddagger}/k_{\text{B}}T} \tag{13}$$

where $F_{\text{kin}}^{\ddagger} = F(n_{\text{kin}}^{\ddagger}) - F(n_{\text{uf}})$ and $n_{\text{kin}}^{\ddagger}$ is the value of n that maximizes the above expression for τ. One may think of $n_{\text{kin}}^{\ddagger}$ as the similarity to the native state where the bottleneck for folding occurs. The set of states with $n = n_{\text{kin}}^{\ddagger}$ acts like the transition state for folding when we consider influences of external agents on rates.

Although Eq. (13) for the folding time has the same form as analogous expressions from traditional transition state theory, there are three important differences. First, the prefactor is $\bar{t}(n_{\text{kin}}^{\ddagger})$, the typical lifetime of an individual microstate at a similarity $n_{\text{kin}}^{\ddagger}$ to the native structure. The corresponding prefactor in absolute rate theory would be an expression involving only fundamental constants. The need for the prefactor based on the lifetime of the microstates stems from the greater complexity of protein folding as compared with the gas phase reactions which absolute rate theory was originally designed to describe. This lifetime strongly depends on the roughness of the surface. Ignoring this fact, we see that the folding is considerably less than the Levinthal estimate, because some of the configurational entropy loss is balanced by the gain in energy as the native structure is approached. The second difference is that $n_{\text{kin}}^{\ddagger}$, the analogue of the transition state in Eq. (13) for the folding time, is determined by maximizing the *entire* folding time expression in this equation. In contrast, in traditional transition state theory, the transition state is a maximum of the free energy, which would here correspond to $n_{\text{th}}^{\ddagger}$. If the average lifetimes $\bar{t}(n)$ were constant, i.e., independent of n, then $n_{\text{kin}}^{\ddagger}$ would equal $n_{\text{th}}^{\ddagger}$. However, in protein folding, we expect the average lifetimes $\bar{t}(n)$ to vary strongly with n, so $n_{\text{kin}}^{\ddagger}$ will *not* always equal $n_{\text{th}}^{\ddagger}$ and the difference can be large and important. More concisely, the position of the *kinetic* folding bottleneck, $n_{\text{kin}}^{\ddagger}$, is not necessarily the same as the position of the *thermodynamic* folding bottleneck, $n_{\text{th}}^{\ddagger}$. Third, whereas in traditional transition state theory the transition state typically is a specific configuration, the transition state in our folding time expression (13) corresponds to an entire band of states in the full configuration space and should not be thought of as a unique configuration. Furthermore, since the potential of mean force of the protein chain is dependent on temperature and solvent conditions, the location of the transition state band will change as the temperature and solvent conditions change. This situation is in marked contrast to the case of small molecules in the gas phase in which the transition state can be thought of as a single structure which is fixed for all reaction conditions.

Notice that the free energy gradient provided by the minimal frustration principle leads to multiple paths approaching this transition state surface as long as the glass transition has not been reached and that this is crucial to overcoming the entropy loss on folding. The expected temperature dependence of the folding time is obtained by combining Eq. (13) for the folding time and Eq. (10) for the average lifetime of a microstate. The result, after taking the logarithms in order to simplify the resultant expressions, is

$$\log\left(\frac{\tau}{t_0}\right) = \frac{F_{\text{kin}}^{\ddagger}}{k_{\text{B}}T} + \frac{\Delta E(n_{\text{kin}}^{\ddagger})^2}{(k_{\text{B}}T)^2}. \tag{14}$$

Notice that if $F_{\text{kin}}^{\ddagger}$ and $\Delta E(n_{\text{kin}}^{\ddagger})$ are assumed to be temperature independent, then Eq. (14) implies that an Arrhenius plot of folding time versus inverse temperature would be curved, and in fact parabolic. Such curved Arrhenius plots are frequently observed in protein folding experiments.[73] Unfortunately, these plots can not be used to derive values for $F_{\text{kin}}^{\ddagger}$ and $\Delta E(n_{\text{kin}}^{\ddagger})$ directly. First, our discussion of microstate lifetimes is rather rough. A more careful treatment shows that the exponent in the expression for the lifetimes (10) must be replaced with a general quadratic in $\Delta E(n)/k_{\text{B}}T$ when the system gets close to the glass transition. Second, and more important, $F_{\text{kin}}^{\ddagger}$ *does* depend on temperature, because the free energies of the unfolded state and the folding bottleneck, the position of the folding bottleneck, the potential of mean force of the protein molecule all change with temperature. Similarly, $\Delta E(n_{\text{kin}}^{\ddagger})$ also depends on temperature. The main point here is that a curved Arrhenius plot of the folding time should be expected as an elementary consequence of energy landscape properties of protein folding.

The glass transition, discussed above for downhill folding, also occurs in systems with bistable free energy functions, in exactly the same way. As before, when the system $T > T_{\text{g}}(n)$ the behavior of a typical

protein can be replaced by the behavior of a statistical ensemble, so Eqs. (13) and (14) for the folding time are valid. For $T < T_g(n)$ the kinetics are dominated by the details of the energy landscape, so Eqs. (13) and (14) for the folding time must be modified. The kinetic behavior in this glassy regime is non-self-averaging, a term we now discuss.

An important feature of protein folding below the glass transition is non-self-averaging behavior. The idea of non-self-averaging is best approached by first discussing its opposite, self-averaging behavior. In simple terms, a self-averaging property is one that depends on the overall composition of an object, rather than its detailed structure. An illustration of this idea is provided by alloys, for example, brass, an alloy of copper and zinc. No order determines whether a particular lattice site is occupied by a copper atom or a zinc atom, so each piece of brass is different on the atomic scale. However, in spite of these differences, all pieces of brass with sensibly the same composition share many properties, for example, hardness, density, electrical conductivity, etc. These properties are called self-averaging because the value of the property, say hardness, of a member of a statistical ensemble, here pieces of brass with the same composition, is almost always nearly equal to the average value of that property over the statistical ensemble.** Notice that self-averaging is a characteristic of the ensemble and the property taken together. Going back to the alloy example, density is a self-averaging property for all pieces of brass with a specified composition, but is not a self-averaging property for all pieces of metal. As a biochemical example, consider the ensemble of amino acid sequences with the same length and amino acid composition as hen lysozyme. The ability to form a collapsed globule with approximately the radius of gyration as a lysozyme molecule is probably a self-averaging property for this ensemble, whereas the ability to fold to a structure that hydrolyzes glycosidic bonds is almost certainly a non-self-averaging property.

The presence or absence of self-averaging of a given property has important practical implications. If a property is self-averaging over some ensemble, then studying that property in one member of the ensemble suffices to learn about the property for all members of the ensemble; if the property is non-self-averaging, then studying that property in one member of the ensemble provides no information about the property for other members of the ensemble. The question of whether or not a given property is self-averaging is also intimately related to the question of whether or not that property is strongly affected by mutations. A mutation will create a new sequence, i.e., a new member of the ensemble. A self-averaging property will behave in the same way in the mutant as in the rest of the members of the ensemble, but a non-self-averaging property will behave differently in each member of the ensemble, including the mutant.

In protein folding there are several different ensembles over which one can average, a few of which we now list, going from the largest, most general ensemble to the smallest, most specific ensemble. First, there is the most general ensemble relevant to protein folding, that of all possible polymers of amino acids. Experiments on random polypeptide sequences explore this ensemble.[35] Next is the set of ensembles of amino acid sequences with fixed amino acid composition. Experiments that investigate random sequences with only a few types of amino acids have studied instances of these ensembles.[36,74] Interestingly, there is some evidence from computer simulations of protein folding that the collapse time for a sequence depends only on its composition; this evidence indicates that collapse time may be a self-averaging property over these ensembles.[90] Finally, there are the ensembles of sequences that fold to a specific structure, e.g., the different lysozyme sequences mentioned in the introduction. These ensembles are studied in research programs that investigate the properties of different mutants of a particular protein.

How do these considerations of the location of a second-order phase transition corresponding to an ideal glass transition *along* the folding coordinate relative to the extrema of the unimodal and bimodal free energy functions affect the kinetics of folding? We see that there are several distinct folding scenarios, which are illustrated in Figure 5 and which we now discuss. As stated above, for a unimodal free energy function, downhill folding, the rate of folding will depend mainly on the lifetimes of the individual microstates. We call this situation a Type 0 scenario. It is analogous to spinodal crystallization studied in materials science.[75] In this case the unfolded state is unstable; from almost any configuration there is a conformational change that will lower the energy with little cost in entropy. Nevertheless, this type of folding transition can still have a folding bottleneck, like the folding transitions in a bimodal free energy function, if the diffusion constant becomes small as in a glass transition. The difference here is that the folding bottleneck in a Type 0 transition will be *entirely* kinetic, so $n^{\ddagger}_{kin}$ will occur at the maximum of $\bar{t}(n)$. In contrast, for a bimodal free energy function the folding barrier will have both kinetic *and* thermodynamic contributions. The Type 0 scenario can further be broken into two subclasses. In the first subclass, which we call Type 0A, the glass transition

**More precisely, consider a statistical ensemble of objects, and some property of the objects in the ensemble. A property is called self-averaging if the fluctuations of the value of that property in the members of the statistical ensemble are small compared to the average value of the property over the ensemble. More detailed discussions of self-averaging can be found in the references on spin glasses that we cited.

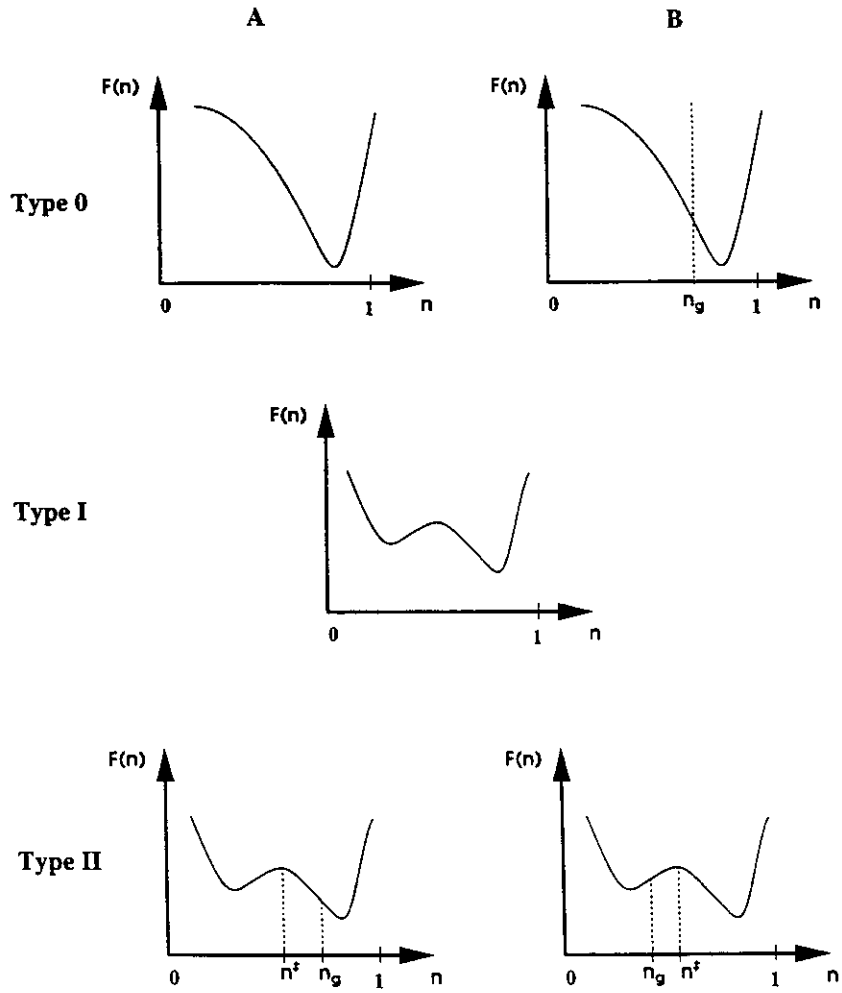


Fig. 5. Schematic illustrations of the folding scenarios discussed in the text. Each sketch shows a qualitative plot of the free energy against the folding coordinate. In the type 0 scenarios shown at the top, the free energy function has only one minimum near the folded state, i.e., $n = 1$. In a type 0A transition, shown at the left, there is no glass transition. In a type 0B transition, shown at the right, at some value of the folding coordinate, n_g, the protein undergoes a glass transition and it exhibits the glassy dynamics described in the text for the remaining of the folding process, $n > n_g$. The type I scenario is shown in the middle of the figure. Here the free energy has two minima, an unfolded one and a folded one, and there is no glass transition during the folding process. The free energy functions in the type II scenarios, shown at the bottom of the figure, also have two minima but the protein undergoes a glass transition during the folding process. In a type IIA scenario, shown at the left, the glass transition occurs after the thermodynamic folding bottleneck at $n_{th}^{\ddagger}$. In a type IIB, shown at the right, $n_{th}^{\ddagger} > n_g$, making the folding protein glassy before the thermodynamic folding bottleneck is reached.

does not occur at any value of n. In this case the folding is fast and dominated by a single rate, the rate of going down the free energy gradient. The kinetics in this regime are self-averaging. In the second subclass, which we call Type 0B, the glass transition occurs before the protein reaches its native state. Then the first part of the folding is a rapid descent down the free energy gradient, as before, but the glass transition intervenes and slows the folding considerably. The overall kinetics is slower and multiexponential because different protein molecules find themselves stuck in a few different microstates after the glass transition, and each of these states will fold at a different rate. Some of the microstate lifetimes can be very long. These long-lived microstates will be observable as kinetic intermediates. The paucity of occupied microstates will lead to discrete pathways as shown schematically in Figure 6. The kinetic behavior is strongly non-self-averaging, so mutations easily change the folding kinetics. Intermediates in one form of the protein are absent in others.

The kinetics of the folding of proteins with bimodal free energy functions fall broadly into two classes. In the first of these, which we call Type I, there is no glass transition at any point in the fold-

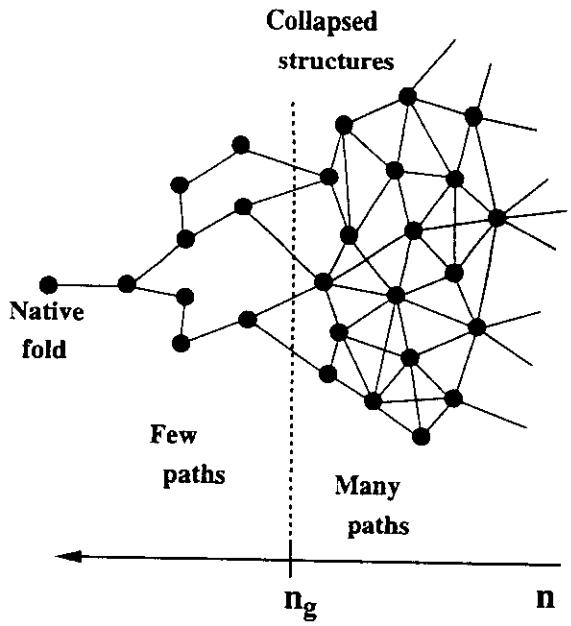


Fig. 6. A schematic representation of the emergence of folding pathways. In this figure the native structure is on the left, so that n increases from right to left. Before the folding protein reaches the glass transition there are many accessible paths between conformations. In this regime each molecule would take a different path as it approached the native structure. After the folding protein goes through the glass transition it has access to only a few paths, so most molecules will take one of a few, or perhaps only one, path to the native structure.

ing, just like the Type 0A folding scenario noted above. Type I scenarios are analogous to nucleation followed by rapid growth.[75] In this case the folding is dominated by a single rate, the folding time being given by Eq. (13). The protein has kinetic access to a representative section of the folding bottleneck, so the rate of folding can be calculated by considering the rate of folding for a statistical ensemble of structures at the bottleneck. In this regime, the protein can take many possible pathways through the bottleneck, so the overall folding time will be independent of the initial unfolded configuration of the protein. The folding kinetics are self-averaging, so mutations will have only small effects on folding rates.[††] In the other class the glass transition occurs at some point in the folding process. We call these folding events Type II. Type II folding processes are analogous to nucleation followed by slow growth: a situation much studied in the metallurgy of alloys.[75] Type II folding scenarios can be broken into two subclasses, depending on where the glass transition occurs relative to the thermodynamic bottleneck location $n_{th}^{\ddagger}$. Recall that $n_{th}^{\ddagger}$ is the location of the maximum of the free energy and need not be the same as the kinetic bottleneck coordinate $n_{kin}^{\ddagger}$ that appears in Eqs. (9) and (13) for the folding time. Thus, for folding at a fixed temperature in a situation where a glass transition occurs, we expect to find two distinct kinetic scenarios, one, which we call Type IIA, occurs when $n_{th}^{\ddagger} < n_g$, and the other, which we call Type IIB, occurs when $n_{th}^{\ddagger} \geq n_g$.

As the roughness of the energy landscape is increased, a glass transition occurs between the folding bottleneck and the final folded state, so that discrete pathways occur after the transition state. We call this situation a Type IIA scenario. In this regime passage through the folding bottleneck will be dominated by a single rate, but there may be some nonexponential behavior, and discrete pathways and kinetic intermediates will be observed in the late stages of folding.

In the Type IIB scenario the protein has already gone through the glass transition when it reaches the maximum of free energy. Since the protein can take only a few pathways after the glass transition, and these pathways can be different enough to lead to wildly different folding times, the overall folding time will strongly depend on which of the few paths to the folded state is taken. Each of these paths will have its own kinetic transition state and the free energies of these states will differ appreciably, i.e., they will not self-average. The importance, and even the meaningfulness, of the *typical* kinetic transition state, $n_{kin}^{\ddagger}$, is diminished considerably in this regime. Therefore we have used the location of the glass transition relative to $n_{th}^{\ddagger}$ rather than $n_{kin}^{\ddagger}$ in defining the difference between the Type IIA and Type IIB scenarios.

THE PHASE DIAGRAM AND PROTEIN FOLDING SCENARIOS

The phase diagram is a powerful tool for understanding protein folding. It reduces much of the discussion about folding scenarios in the previous section to a single, clear, coherent picture which is useful for thinking about and planning experiments.

The simplified viewpoint of protein folding, using the energy landscape framework that we discussed in the last section, can be used to classify different mechanisms of protein folding in the laboratory and in computer simulations. The analysis discussed above uses only a single parameter, n, to characterize the difference between the native structure and the unfolded structures. In fact, native proteins differ from unfolded ones in several ways, so this requires the introduction of several different similarity measures in thinking about folding processes. It is important, however, that the number of additional parameters is relatively small, thus giving a

[††]To be more precise, rates of individual events depend on the exponentials of free energies. Above the glass transition these free energies should all self-average and the significant rates will have a log-normal distribution. A few factors of two change in the rate is not considered significant here. In the glassy phase a much wider distribution of the logarithm of the rate is anticipated, as pointed out by Bryngelson and Wolynes.[5]

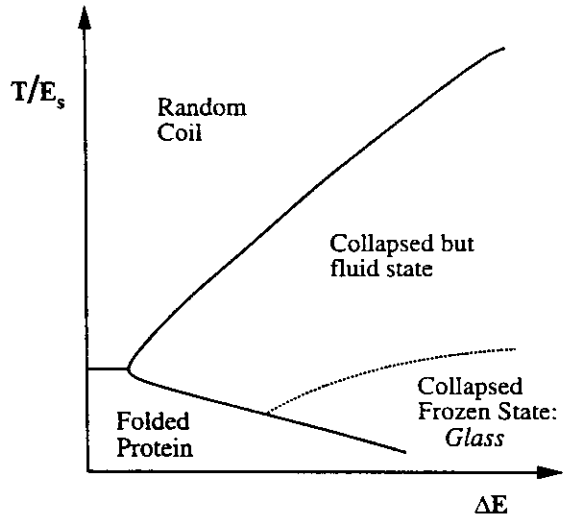


Fig. 7. Phase diagram for a folding protein. The horizontal axis is the energy landscape roughness parameter, ΔE, discussed in the text. The vertical axis is the temperature divided by the stability gap E_s. The stability gap is the energy gap between the set of states with substantial structural similarity to the native state and the lowest of the states with little structural similarity to the native state. The collapse transition and the (first-order) folding transition are represented by solid lines and the (second order) glass transition is represented by a dashed line. In comparing this phase diagram with experimental phase diagrams, one must bear in mind that both ΔE and E_s are temperature dependent because of the hydrophobic force. In addition, the collapse transition depends on the average strength of the hydrophobic force, and this is both temperature and pressure dependent. The average strength of the hydrophobic force could be considered as a third dimension in the phase diagram.

reduced description of the folding process. Indeed, many of the discussions of folding pathways have concentrated on these additional similarity measures or order parameters. Thus in many pictures of protein folding, e.g., the framework model,[76] one gives considerable emphasis to the initial formation of secondary structures. In other scenarios, the collapse and formation of secondary structures are considered to be separate events.[6] Additionally, proteins may consist of subdomains for which we may discuss the tertiary structure formation separately. This is particularly important in hierarchical pictures of protein folding.[77] With each of these similarity measures we can ask the way in which the formation of order is related to the roughness of the energy landscape and whether the transition occurs through many pathways or through a small number of distinct pathways. It is helpful to consider a phase diagram like the one illustrated in Figure 7.

In this phase diagram, we plot the possible equilibrium states of a protein as a function of temperature and roughness of energy landscape. The phase diagram contains a region of random coil, a collapsed phase, a folded region with transition lines between these places, as well as a dotted line indicating the presence of a frozen glassy state.

A given protein will exist at equilibrium somewhere in this phase diagram, thus the diagram tells us the final state which we would obtain in an experiment. The folding process begins by starting in a configuration characterized by one of the regions on this diagram, but is carried out at a temperature such that the folded protein is the lowest free energy state. The roughness of the energy landscapes is important in determining the equilibrium phase but plays a bigger role in the kinetics of the folding process as described before. In the lefthand part of the diagram, folding will occur by a Type I mechanism in which discrete pathways are not observed. As the roughness is increased, the folding can occur by a Type IIA mechanism in which discrete pathways occur after the transition state. As the roughness of the energy landscape increases more, and the equilibrium glass transition occurs before the transition state is reached, the folding occurs through a Type IIB mechanism in which discrete pathways are observed and misfolded states play a role in the dynamics. Structurally unique thermodynamic transition states can occur only if $T < T(n_{\text{th}}^{\ddagger})$, i.e., if the folding is Type IIB, because that is the only case where there are order one accessible paths through the folding bottleneck. In all other folding scenarios, there are many accessible paths through the folding bottleneck, hence many possible transition states.

The temperature at which the folding experiment takes place also plays an important role in whether a Type 0, Type I, or Type II scenario for folding is observed. At low temperatures, (relative to the roughness energy scale) one expects to see nonexponential kinetics characterizing a Type IIB scenario. On the other hand, at higher temperature, at the midpoint of the folding transition, one expects Type I or IIA mechanisms to be more prevalent. Since ruggedness appears only when contacts are made, when there is little frustration, as well as little average driving force toward hydrophobic collapse, a Type I mechanism is most probable. This is very close to the framework model[76] or diffusion collision-picture[78–82] that was so often thought to describe protein folding. In the original versions of such models, only correct structures are formed initially and these can dock to form completed structures. Such a highly unfrustrated situation seems to be uncommon and certainly does not occur in the computer simulations of protein-like models.

Good folding sequences are ones that have a strong free energy gradient leading to the ground state structure. To achieve this they must separate in energy the native conformation and those conformations that are structurally similar to the native conformation from the bulk of most of the other conformations with no structural similarity to the native conformation. Goldstein et al. have shown that this qualitative criterion is equivalent to finding sequences that maximize T_f/T_g for a suitable simpli-

fication of the Bryngelson–Wolynes model.[83] Notice that the energy gap that is being maximized when T_f/T_g is maximized is *not* the energy gap between any two specific states, but rather the gap between the *set* of states with substantial structural similarity to the native state and the lowest of the *set* of states with little structural similarity to the native state. We call this gap the "stability gap" (E_s). The stability gap should not be confused with the energy gap between the native configuration and the configuration with the next highest energy. This state will usually be native-like itself. There are too many fluctuations in the folded state for this two-configuration energy gap to have any significance for protein folding or stability.[17,18,58,85] In fact, Frauenfelder and collaborators have interpreted the results of their experiments on *folded* proteins in terms of a hierarchy of "substates." These substates correspond to slightly different structures found in the population of folded proteins.[18] Evidence for the highest level of this hierarchy has been seen in protein folding simulations.[59] Unfortunately, this issue of energy gaps has been clouded by lattice simulations that have studied the energy spectrum of only the maximally compact states.[71, 86] Since the maximally compact states are a small fraction of all possible states and since they are often not dynamically connected (in the sense described in the second section),[15, 40] the interpretation of the results of these simulations requires more subtlety than has been found in the literature so far. Notice, however, that since local excitations from a maximally compact state are not themselves maximally compact, the energy gap between the native state and the next lowest energy maximally compact state is often correlated with the stability gap. Thus, the results of these simulations can be interpreted as a confirmation of the older and more general idea that the sequences with large stability gaps fold quickly at the equilibrium folding temperature.[4–6,83,84]

In the Bryngelson–Wolynes energy landscape the stability gap is a tautological consequence of the greater degree of stability of native-like interactions demanded by the principle of minimal frustration. Goldstein et al. calculated a set of parameters that maximized T_f/T_g for the model used in their protein structure prediction algorithm, and found that these parameters gave excellent results for practical structure prediction, in accord with the predictions of the theory. In addition, molecular dynamics calculations using associative memory Hamiltonians optimized in this way reliably gave native-like structures.[83,84] These results provide independent evidence that sequences that satisfy this criterion (of having a large stability gap) should be good folding sequences. This work also is a good illustration of the power of using energy landscape ideas to help solve practical protein folding problems. We also mention that the stability gap idea has been used by Wodak and co-workers to predict persistent secondary structures in small peptides relevant to early folding events.[87]

The phase diagram, of course, becomes more complex as additional order parameters or similarity measures are used to characterize the folded states. The phase diagram is a useful way of thinking about any folding process because it allows us to consider the couplings between the various order parameters as well. For instance, as one sees in the computer simulations, one can first have a collapse which is ascribed by a single-order parameter, radius of gyration, followed later from this collapsed phase by a transition to a unique folded protein structure.[40,88–90] The coupling between these two parameters is crucial in obtaining that sort of description. The so-called molten globule intermediates which are often an ensemble of individual configurations really should be described by these additional order parameters.[68]

ENERGY LANDSCAPE ANALYSIS OF FOLDING SIMULATIONS

Simulations of simple protein-like lattice models provide an ideal ground to illustrate the energy landscape ideas. Lattice models have a venerable history.[50–52,91–100] There is widespread agreement that they capture some of the underlying physics of protein folding. There are also excellent reviews that discuss lattice simulations in the context of the general problem of understanding protein folding.[63,101,102] Many groups have interpreted their simulation results using some of the qualitative and semiquantitative ideas of energy landscape analysis, finding features in agreement with the overall picture that we have just discussed.[40–42,58,59,71,103,104] Here we illustrate this kind of discussion by focusing on some recent results of Socci and Onuchic which find evidence for specific features arising from energy landscape analysis.[90] In addition, these simulations provide an excellent example of the kind of quantitative analysis which should be carried out for real experimental data. We will use simplified quantitative relations that can be deduced from the energy landscape analysis. This sort of quantitative analysis should also be carried out for laboratory experiments, but in the laboratory the temperature dependence of the various free energy contributions must also be included explicitly for a fully convincing analysis. Simulations based on reduced models avoid these issues since the energy function is itself not temperature dependent.

The simulations were performed on polymers that were 27 monomers long which have maximally compact states of $3 \times 3 \times 3$ cubes. Because the configurations on the $3 \times 3 \times 3$ cube can be completely enumerated in a reasonable amount of computer time, the energy landscape among the maximally compact states can be explored in great detail. This 27 mono-

TABLE I. Various Sequences Used*

Run	Sequence	E_{min}	τ_{min}	T_g	T_f
002	ABABBBBBABBABABAAABBAAAAAAB	−84	2.0×10^7	1.00	1.285 (15)
004	AABAABAABBABAAABABBABABABBB	−84	1.6×10^7	0.96	1.26 (1)
005	AABAABAABBABBAABABBABABABBB	−82	2.3×10^7	0.98	1.15 (2)
006	AABABBABAABBABAAAABABAABBBB	−80	5.2×10^7	1.07	0.95 (6)
007	ABBABBABABABAABABABABBBABAA	−80	9.3×10^7	1.09	0.93 (5)
013	ABBBABBABAABBBAAABBABAABABA	−76	9.7×10^7	1.01	0.83 (5)

*The last four (005, 006, 007, 013) were generated at random. Sequence 002 is from ref. 132. Sequence 004 is a single monomer mutation of 005 ($B_{13} \rightarrow A$). Both 002 and 004 have the lowest energies possible for the potential used and have native states that are completely unfrustrated, i.e., very native contact is individually stabilizing. τ_{min} is the fastest folding time for each sequence. T_g is the glass transition temperature (calculated with a $\tau_{max} = 1.08 \times 10^9$). T_f is the folding temperature calculated using the Monte Carlo histogram method. The numbers in parentheses indicate the uncertainty of the last digit.

mer cubic simulation has been a paradigm of study in this field because of this feature.[54,55,71,86,105] The simulations of Socci and Onuchic contain two monomer types. Pairs of monomers that were nearest-neighbors on the lattice but not connected along the chain contributed an interaction energy to the potential. The potential for the two monomer code was −3 for contacts between monomers of the same type and −1 for contacts between different types. The folded configuration was taken to be the maximally compact configuration with the lowest energy.

The characteristic energy scales and temperatures for different sequences are easily obtained for these models. The folding temperature, T_f, may be defined in the usual way as the temperature at which population in the folded configuration is equal to the populations in all other configurations. These populations can be obtained by a Monte Carlo sampling procedure for each of the sequences. The folding temperatures correlate rather well with the energy of the folded configuration. This is shown in Table I. Figure 8 shows the equilibrium folding curves for these sequences.

A kinetic glass transition temperature can be defined without appealing explicitly to the energy landscape analysis. Just as in a laboratory, a kinetic glass transition temperature is defined by asking where a characteristic timescale in the problem exceeds some large value. In the simulations the maximum running time was $\tau_{max} = 1.08 \times 10^9$ Monte Carlo steps. This number was chosen because it was significantly longer than the folding times over a broad range of temperatures. It would be appropriate to define the characteristic time through the typical time for a large-scale rearrangement. However, it is simpler here to use the folding time itself as a timescale. A kinetic glass transition temperature, T_g, then is defined by the criterion $\tau_f(T_g)$ is $(\tau_{max} + \tau_{min})/2$ where $\tau_f(T)$ is the folding time at temperature T. As you can see from Table I, this transition is nearly self-averaging, that is, it depends very little on the particular sequence which is studied, and is roughly 1.0.

According to the energy landscape analysis this

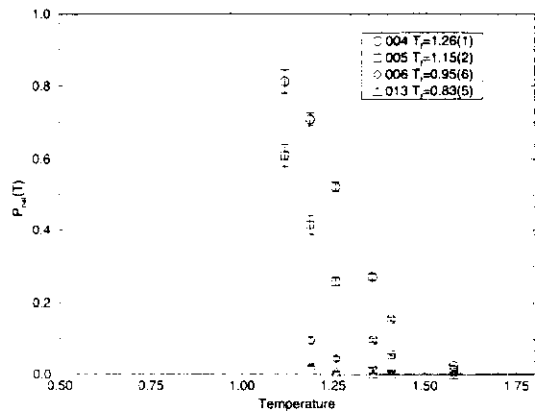


Fig. 8. Folding curves for four of the sequences used in the simulation. The probability of a sequence occupying the native structure is plotted on the vertical axis versus temperature on the horizontal axis. The folding temperature, T_f, is defined as the temperature where $P_{nat}(T_f) = 0.5$, i.e., the probability of the occupancy of the native structure is one-half. The numbers in parentheses indicate the uncertainty in the last digit.

kinetic glass transition is most strongly influenced by the thermodynamic glass transition. The simulations bear out this expectation. Changing the fiducial cut-off time by a factor of 8 causes only a 10% change in the kinetic T_g. Similarly, small changes to the algorithm for selecting the moves have a small effect on T_g.[‡‡]

The thermodynamic glass transition of the BW analysis depends on the entropy and roughness energy scale of the compact states. This thermodynamic T_g is also a self-averaging quantity. Using only the maximally compact cube states, one obtains $T_g \approx 1.17$. This estimate of T_g is likely an upper bound, since semicompact states also contribute to the entropy. At the same time, kinetic constraints could create additional restrictions on this connectivity. These effects seem to cancel, so the kinetic and thermodynamic glass transitions are rather

[‡‡]Only if the number of crankshaft moves is reduced to less than 10% of the corner moves is there any very dramatic change in T_g.

close and one can take them both to be approximately 1 in analyzing the figures.

Figure 9 shows a plot of the folding time, that is, the time it takes for a random unfolded initial condition to reach the native structure, for different temperatures.[§§] Since the 27 monomer length heteropolymer is so small, it is possible to analyze folding both above and below T_f for quite a range of temperatures. Above T_f the folding process is essentially an uphill one but with a modest slope. The first noticeable feature about the folding data is that they are strongly sequence dependent at intermediate temperatures. In the simulations folding times greater than a maximum of τ_{max} were assigned the folding time τ_{max}. This is the origin of the saturation at the high and low temperature ends of these curves. At high temperature the folding is slow because it is so strongly uphill entropically. At low temperatures the folding is slow because of the roughness of the energy landscapes for all of these sequences. Another characteristic feature, however, is that the folding time at intermediate temperatures is most strongly correlated with the stability of the folded state for each of the sequences. The fastest folding sequence has the highest folding temperature, while the slowest has the lowest folding temperature. Indeed, the slowest folding sequence has a folding temperature less than the glass transition temperature.

Also plotted in Figure 9 are two different collapse times for the same sequences. The lower curve is the time that it takes the sequence to encounter, for the first time, a structure with 25 contacts. The middle curve is the time needed by the protein to achieve any maximally compact 28 contact cube. The remarkable qualitative feature of these collapse time curves is that at the moderate to high temperatures where the folding times vary greatly, all of the sequences have essentially the same collapse times. In this temperature range collapse is a self-averaging process that depends primarily on the average composition of the protein molecules. Another remarkable feature, however, is that the collapse time begins to fluctuate greatly between different sequences at and below the kinetic glass transition temperature. The energy landscape analysis suggests that individual transition times between states fluctuate greatly below T_g, and this is reflected in the collapse process. The distribution of folding times becomes broader as you approach T_g, reflecting the emergence of a multiexponential collapse process. We note that Flanagan et al. have observed sequence-dependent collapse in staphylococcal nuclease.[106] This suggests the phase observed is near its glass transition.

A rough quantitative understanding of these data for folding and collapse comes from energy land-

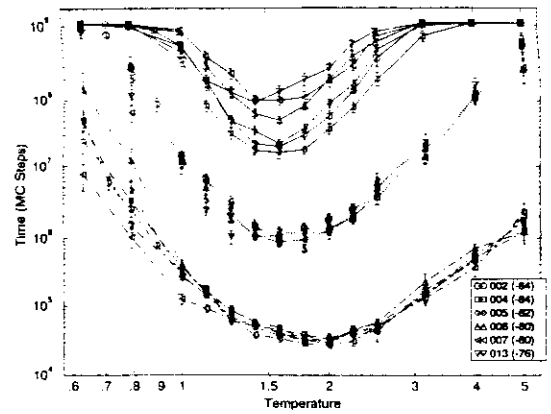


Fig. 9. Plots of important times (in Monte Carlo steps) against temperature for the sequences used in our simulations. The top curves are the folding times τ_f (the number of steps required to reach the native structure for the first time). The saturation at the wings of the curves occurs because runs were stopped at a maximum time of 1.08×10^9 Monte Carlos steps. The other curves are plots of collapse times. The middle curves are the times required for the sequences to reach a conformation with 28 contacts for the first time. Similarly, the bottom curve is the time required to reach a conformation with 25 contacts for the first time. Notice that there is a much greater time spread in the folding curves than in the two collapse curves.

scape analysis. The availability of both folding and collapse results allows us to roughly separate features connected with the glassy dynamics from the thermodynamic changes that also result from rough energy landscapes. The first important observation is that both folding and collapse times give parabolic Arrhenius plots, just as most experimental data do for the forward and reverse rate of folding.[73] In the laboratory this curvature is usually ascribed to the thermodynamic dependence of the effective interactions, the difference of heat capacity between the folded and unfolded states arising from the hydrophobic effect. Since the force laws in the simulation are taken to be *independent* of temperature, the temperature dependence of the hydrophobic effect is not at all involved in the simulation data. The simulation of Miller et al. also effectively finds a curved Arrhenius plot.[107] A simple analysis can be carried out by assuming that the location of the folding bottleneck, $n^{\ddagger}_{kin}$, is independent of temperature. Roughly speaking then, the folding time will be given by Eq. (13) with the energy barrier $F^{\ddagger}_{kin}$ given by the difference in free energy between the folding bottleneck states and the free energy of the bottom of the unfolded free energy minimum, i.e., the lowest free energy unfolded states. This involves motion on the free energy gradient for the reaction coordinate based on the number of correct contacts. At this level of analysis, the collapse time can be treated in a similar way using the total number of contacts of any kind as a reaction coordinate. In the temperature range of 1.0 to 2.25 (the reason for considering

[§§]Technically, these times are mean-first-passage times.

this temperature range will become clear below) the time required for collapse to configurations with 25 contacts varies by a factor of less than 4, indicating that there is little, if any, free energy barrier to collapse. Therefore, collapse is essentially downhill in free energy and behaves like a Type 0A scenario. The dynamic reorganization timescale will become longer as the protein becomes more compact because excluded volume has a stronger effect on dynamics in compact states. Therefore, in the generalized transition state approximation of the section on Quantitative Aspects of the Kinetics of a Folding Protein, the collapse time will be given by Eq. (9) for the time for a downhill process,

$$\tau_{\text{collapse}} = \bar{t}_{\text{collapse}} \tag{15}$$

where $\bar{t}_{\text{collapse}}$ is the typical lifetime of an individual microstate in a random collapsed state. For the purposes of calculating the barrier height, $F^{\ddagger}_{\text{kin}}$, we set the free energy of the bottom of the unfolded free energy minimum equal to the free energy of the collapsed states. Then the folding time involves the free energy difference of the folded and compact configurations. Another way of obtaining a folding time that depends on this free energy difference is to consider folding to be a three state unimolecular reaction, *random coil* $\rightleftharpoons$ *collapsed* $\rightarrow$ *folded*, where the second step, *collapsed* $\rightarrow$ *folded* is rate-limiting. The data are consistent with such a reaction scheme.

We can eliminate the purely dynamic factors by taking the ratio of the folding to the collapse time and assuming that $\bar{t}(n^{\ddagger}_{\text{kin}}) \approx \bar{t}_{\text{collapse}}$. Then using Eq. (13) for the folding and collapse times and using Eq. (8) for the free energy predicts that a plot of the logarithm of the ratio of the folding to the collapse times is parabolic,

$$\log\left(\frac{\tau}{\tau_{\text{collapse}}}\right) = -[S_0(n^{\ddagger}_{\text{kin}}) - S_{0,\text{collapse}}] + \frac{[\bar{E}(n^{\ddagger}_{\text{kin}}) - \bar{E}_{\text{collapse}}]}{k_B T} - \frac{[\Delta E(n^{\ddagger}_{\text{kin}})^2 - \Delta E^2_{\text{collapse}}]}{2(k_B T)^2}, \tag{16}$$

where the subscript collapse indicates that the quantity is evaluated in a random collapsed state. The log of this ratio is plotted versus $1/T$ in Figure 10. We show here the data only between temperatures 1.0 and 2.25 because outside this range the folding times exceed the time used as a cut-off in the simulations. These curves can be fit very adequately with parabolas. The coefficients of the parabolas are shown in Table II. In the fit all of the constant terms are positive and all of the linear (in $1/T$) terms are negative, which imply the inequalities $S_0(n^{\ddagger}_{\text{kin}}) < S_{0,\text{collapse}}$ and $\bar{E}(n^{\ddagger}_{\text{kin}}) < \bar{E}_{\text{collapse}}$. Both of these inequalities are consistent with the bottleneck for

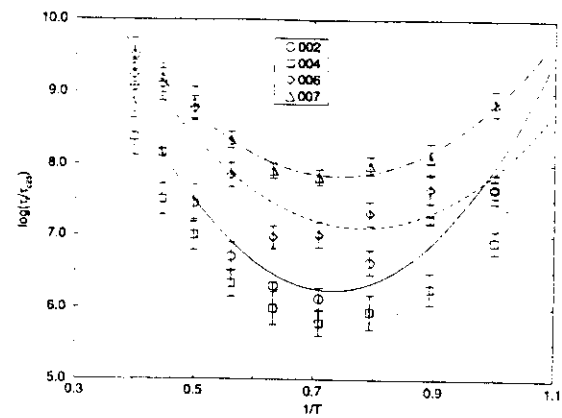


Fig. 10. The logarithm of the ratio of the folding time to the time for collapse to 25 contacts against the inverse temperature. The lines are parabola fits to the data. The coefficients of these parabolas are shown in Table II.

folding occurring after the collapse, in agreement with both intuition and the simulation data. The curvature reflects the value of the roughness of the energy landscape of the collapsed configurations. This analysis shows that for a rough energy landscape, the heat capacity of the collapsed configurations arises from fluctuations in structure and corresponding energy differences between collapsed configurations. The linear term in $1/T$ reflects primarily the enthalpic part of the activation free energy for achieving a transition state. It should be strongly correlated with the stability gap.

One can also check the theory by using independently derived information about the simulation model to make order-of-magnitude estimates of the sizes of the coefficients in the parabola fits. The constant term is the difference of the configurational entropies of the collapsed states and the folding transition bottleneck states. The number of states with 25 contacts has been estimated to be 10^9, yielding a configurational entropy of $9 \log 10 \approx 21$. The configurational entropy of the folding bottleneck states is more difficult to estimate, but it is clearly less than that of the collapsed states. Therefore, the constant coefficient is expected to be of order 10, i.e., between ≈ 3 and ≈ 30. This expectation is very well confirmed by Table II, where the constant coefficients are seen to lie between 16 and 19. The coefficient of the $1/T$ term is the difference between the average energy of the folding bottleneck states and the collapsed states. We have defined a collapsed state to be a state with 25 contacts and the average contact energy in our model is -2, therefore, the average energy of a collapsed state is -50. The average energy of a folding bottleneck state must be greater than the energy of the native state, which is -84 for sequences 002 and 004 and -80 for sequences 006 and 007 (see Table I). Therefore, we

TABLE II. The Coefficients of the Parabolic Fits, $\log(\tau/\tau_{collapse}) = A + B/T + C/T^2$, to the Data Shown in Figure 10*

Run	$A = S_0(n_{kin}^{\ddagger}) - S_{0,collapse}$	$B = \bar{E}(n_{kin}^{\ddagger}) - \bar{E}_{collapse}$	$C = (1/2)[\Delta E(n_{kin}^{\ddagger})^2 - \Delta E^2_{collapse}]$
002	19.0	−34.9	23.8
004	17.0	−30.1	20.1
006	16.9	−25.2	16.1
007	16.1	−22.2	15.0

*The sequence numbers refer to the sequences displayed in Table I. The column headings also show the physical chemical interpretations of the coefficients given in Eq. (16) in the text.

expect the coefficient of the $1/T$ term to lie between 0 and −34 for the first two sequences and to lie between 0 and −30 for the later two sequences. Table II shows the coefficients to lie within these bounds, within reasonable error estimates. The coefficient of the $1/T^2$ term is one-half times the difference between the roughnesses of the collapsed states and the bottleneck states. Each interaction energy in the model differs from the average interaction energy by +1 or −1, so the roughness of the set of random collapsed states with 25 contacts is $\Delta E = 25$. The roughness of the bottleneck states is smaller than this number, but difficult to estimate. Thus, we expect the coefficient of the $1/T^2$ term to be somewhat less than 12.5. The values for this coefficient range from 14 to 24, as shown in Table II. This estimate is not as good as the previous ones, but it does give the right sign and order-of-magnitude, which is the best that can be expected from such an approximate theory and such simple estimates.

A quantitative relationship between protein folding kinetics and the thermodynamic stability of the native state can be obtained with linear free energy relationships.[70,108–110] In the past these relations have been applied to the interpretation of data from site-directed mutagenesis experiments.[111–114] They are also the mainstay of the analysis of many other biochemical reactions.[115–117] In this analysis the differences in the free energies of the transition states, folded states, and unfolded states for two different sequences obey the linear relation

$$\delta F(n_{kin}^{\ddagger}) = \alpha \delta F(1) - (1 - \alpha)\delta F(0) \qquad (17)$$

The transfer coefficient α is a measure of the resemblance of the transition state to the folded state. The value of α is easily obtained from the data. If we make the obvious assumption that the dynamic factors are approximately the same for the different sequences, then Eq. (17) implies that a plot of the logarithm of the folding rate against the logarithm of the equilibrium constant for folding will be a straight line with a slope of α.[118] When we plot the logarithm of the folding rate versus the logarithm of the equilibrium constant for different sequences, we see such a nice linear free energy relationship, shown in Figure 11. At the temperature $T = 1.0$ the folding time seems nearly independent of the driving force, while the driving force is entirely reflected in the unfolding rate. Thus folding here is nearly entirely "downhill," a Type 0 scenario. (The large fluctuations suggest a Type 0B.) At $T = 1.26$ there is a clear nucleation barrier, but it is small. The transfer coefficient of $\alpha = 0.1$ suggests a rather early transition state, i.e., at this temperature the bottleneck configurations are collapsed but have little native structure. The further increase of α at higher T reflects a later transition state as the entropy terms become more important. This shows that the transitions are only weakly Type I and essentially Type 0 under these thermodynamic conditions. The success of this analysis is remarkable because the native structures corresponding with the sequence are not strongly related to each other unlike the situation in site-directed mutagenesis experiments.

An Arrhenius plot of the unfolding time versus $1/T$ is shown in Figure 12. This curve shows the dynamic effects as T_g is approached. There is a clear change in the behavior of the activation energy for unfolding near T_g where the curve starts to level off. This behavior reflects the change in dynamics at T_g suggested by the energy landscape analysis. The loss of dynamic flexibility caused by the entropy crisis leads to dynamic reorganization times limited by the entropy of search and the activation energy of the elementary step [see Eq. (11)]. This analysis of the computer simulation shows many of the ways in which data can be reduced when the thermodynamic dependence of the underlying forces is understood.

ENERGY LANDSCAPE ANALYSIS AND FOLDING EXPERIMENTS

We now turn to the analysis of some particular proteins that have been studied extensively in the laboratory, lysozyme, chymotrypsin inhibitor, and cytochrome c. Despite the significant work already done on these systems, we believe that there are insufficient data to uniquely classify the mechanisms of folding via our energy landscape framework. However, it is possible to use the existing data to give a flavor of how these ideas can be used in laboratory situations. As we have seen in our discussion of the computer simulations, many qualitative features of experiments, such as curved Arrhenius plots, can be obtained from the energy landscape scenario, and can even be quantified if the underlying driving forces are understood. A considerable difficulty in the experimental studies is that these driving forces are temperature dependent.[13] It is, however, important to realize that we can separately change the driving force by such devices as the use of denaturant or mutation and separate this effect from those effects which are directly due to the ruggedness of the energy landscape due to thermal

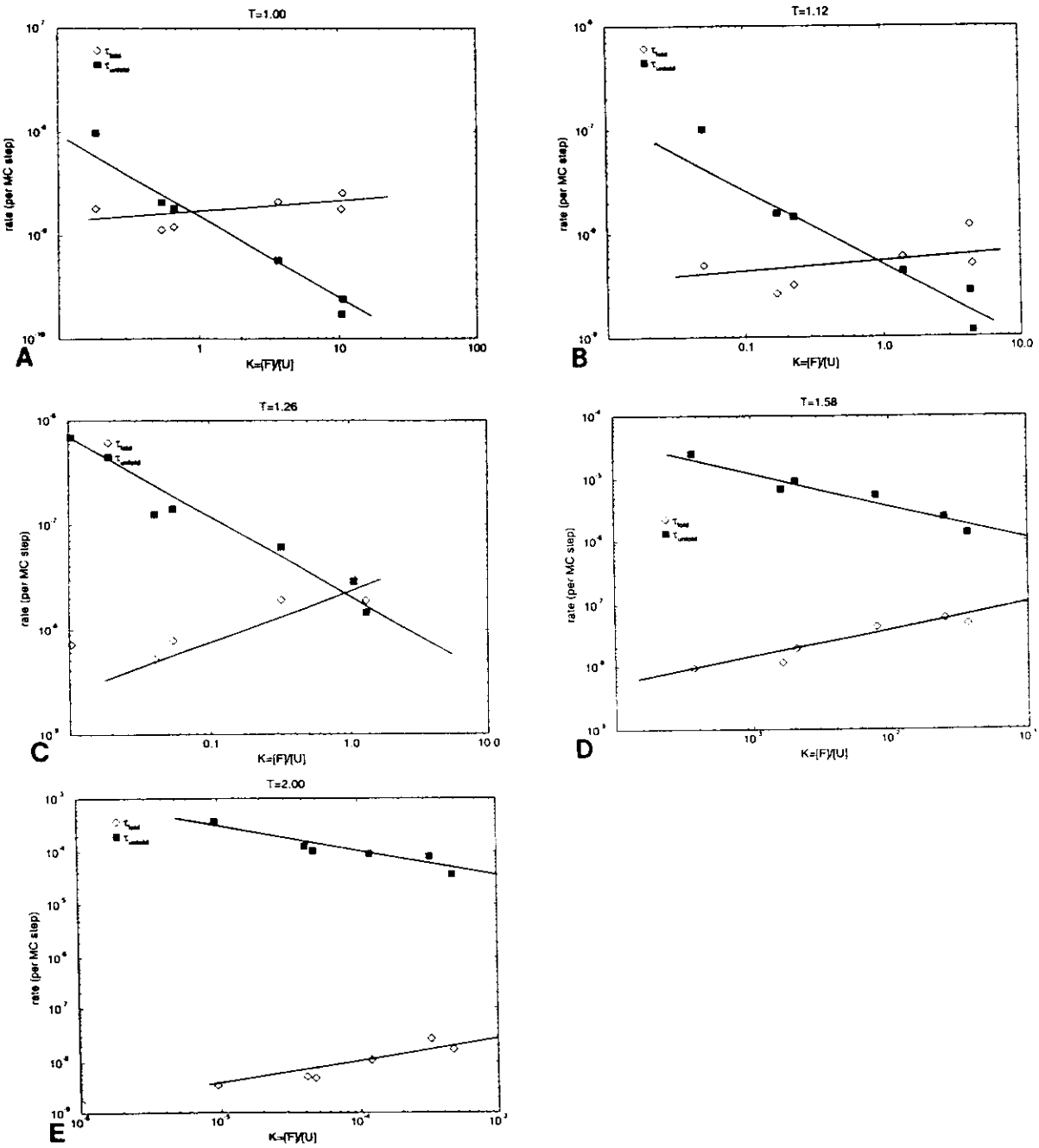


Fig. 11. Plots for the linear free energy relationship analysis. Each plot shows the folding rate against the folding equilibrium constant for each of the six sequences studied here. On the horizontal axis [*F*] represents the probability of the native structure being occupied and [*U*] represents the probability of a non-native structure being occupied. **(A)** A linear free energy plot for temperature $T = 1.0$, that is, at about the glass transition temperature for these sequences. The rest of the plots **(B–E)** are for temperatures above the glass transition temperatures.

energies. Ruggedness is a more nearly self-averaging quantity. A further analysis of this type for specific systems will, we hope, be made soon.***

***At this point the reader may wish to review the folding scenarios discussed in Figure 5.

In some ways, the simplest experimental situation occurs for those proteins and conditions which exhibit a Type I folding mechanism. The kinetics in such system should be simple exponential. These systems have moderate driving forces and are studied in the near equilibrium range near the midpoint

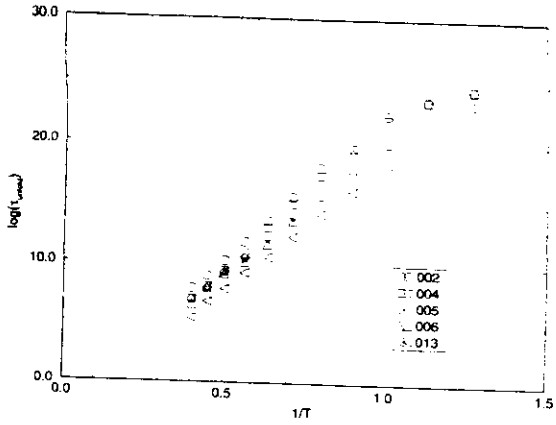


Fig. 12. An Arrhenius plot of the unfolding time against the inverse temperature for five of the sequences. The unfolding time was calculated by multiplying the folding time by the ratio of the folded population to the unfolded population at a given temperature. Consequently, there is the same saturation effect at low temperature as in Figure 9 caused by the finite simulation time.

of the transition curve. One feature favoring a Type I transition as opposed to a Type II transition is the avoidance of premature collapse. When collapse occurs corresponding changes in the ruggedness of the energy landscape can arise and play a role. Apparently Type I behavior occurs upon the cold denaturation of lysozyme as studied by Chen and Schellman.[119,120] It is substantially a uniexponential process.

The folding of chymotrypsin Inhibitor 2, an 83 residue monomeric protein with no disulfide bonds, has been studied by Jackson and Fersht.[121,122] In many respects their experiments resemble a Type I scenario. Jackson and Fersht used fluorescence measurements and scanning microcalorimetry to study the refolding of this protein. The equilibrium denaturation experiments found strong evidence for a simple two-state transition without intermediates. The kinetic measurements, however, reveal three phases, but it is clear that these are due to the five proline residues in the molecule, of which at least four are in the *trans* state in the crystal structure. Seventy-seven percent of the protein molecules fold with a time constant of 0.02 s and the two observable slow phases have time constants of 43 and 500 s. The slow phases are catalyzed by peptidyl-prolyl isomerase, which catalyzes proline isomerization. The fast phase is not affected by this enzyme. The protein molecules that start with all the prolines in the *trans* configuration have very nearly exponential kinetics on the timescale studied.

In the energy landscape view, proline isomerization appears as a high ridge separating the configurations with a *cis* isomer from those with a *trans* isomer.[123,124] One such ridge appears for each proline in the protein. Each of these separate parts of the configuration space can be analyzed with the simple energy landscape concepts that we have already discussed. Thus, the mere observation of multiexponentiality is not enough to imply that these systems obey Type II kinetics in which a glass transition is present. These ridges in the energy landscape come from the simple effect of single amino acid residues, whereas the glass transition comes from the composite effect of all the amino acid residues in the protein.

An example of apparent Type II behavior is provided by hen lysozyme at its high temperature denaturation transition. The evidence for Type II behavior of lysozyme at this transition is largely based on the CD measurements and pulsed hydrogen-exchange labeling carried out by Radford et al.[125] These studies suggest multiexponential behavior for the protection of the amide hydrogens, which Radford et al. have interpreted as due to the existence of multiple parallel folding pathways. The Type II nature of this transition apparently occurs because of the possibility of early collapse. In addition, misfolding is apparently present since the CD shows, after the first 100 ms, considerably more α-helix present than is present in the native state. Thus, in this situation, the folding protein adopts a locally favorable conformation which must be partly unfolded to get into the globally favored native state. The initial strong local tendency toward helix formation is giving rise to frustration in the technical sense of competing interactions discussed earlier in this paper. The Type II behavior suggests that the roughness of the energy landscape for lysozyme is actually larger, compared to k_BT, at the high temperatures than at the low temperatures, apparently due to the temperature dependence of the hydrophobic forces.

Cytochrome *c*, with its heme constraints, apparently has little roughness to its energy landscape compared to the free energy gradient. The heme is covalently bound to the protein chain and after the iron coordination sphere is completed, folding of different parts of the protein occurs rather rapidly. On the other hand, the heme group can also be misligated by some of the amino acids in the protein and this misligation can be detected spectroscopically. The misligated population cannot follow the free energy gradient all the way to the native structure so the presence of the heme also facilitates the study of the different misfolded structures present in an ensemble of folding proteins. Sosnick et al. have studied the folding of cytochrome *c* under conditions where the misligation does not occur.[126] They found that about 50–70% of the molecules in this population acquired native secondary and tertiary structure with a time constant of approximately 15 ms. They estimated, from fluorescence quenching, the time constant for collapse to be approximately 12 ms, that is, of the same order as the folding time. These experiments suggest that cytochrome *c* folding is Type 0 under these conditions though it is difficult to assign it to Type 0A or Type 0B with the

data from these experiments. Experiments on cytochrome c folding provide a good illustration of how the folding of a particular protein can vary qualitatively as the conditions of the folding experiment vary. For example, when cytochrome c is refolded at pH 6.2, the folding is multiexponential and takes on the order of seconds. Sosnick et al. have also shown that the slow folding at pH 6.2 is due to the formation of misfolded, collapsed structure, rather than the specific misligation of the heme, in agreement with the picture of a glassy phase presented here.

CONCLUSION

The energy landscape picture allows us to combine various disparate ideas about the nature of biomolecular self-organization in protein folding. The energy landscape picture can accommodate multiple parallel path scenarios, as well as unique, sequence-dependent pathways for protein folding. The crucial concept in understanding particular experimental and computer simulation situations is to organize the kinetics of the problem through the consideration of a phase diagram and to study the dynamics of the crucial order parameters for folding which distinguish folded states from unfolded ones. In a generic energy landscape picture, several different phase transitions occur and are coupled. At the very minimum, one must consider the two purely thermodynamic transitions of folding and of collapse. The collapse transition temperature depends upon both the overall tendency for self-association and also on the ruggedness of the energy landscape. Above the glass transition collapse is a largely self-averaging process; that is, it depends on the overall composition of the sequence and on little else. The folding transition, on the other hand, is always sensitive to the details of the sequence. In addition to these conventional, understood phases, a rough energy landscape exhibits a glass transition which occurs near a thermodynamic glass transition temperature, T_g. This temperature is also a self-averaging property of different sequences of similar composition.

Different scenarios for protein folding mechanisms occur, depending on the relationship of these various temperatures and the conditions under which the experiment is carried out. The simplest situation to understand occurs when there is a moderate driving force toward the folded state. Near the midpoint of the denaturation curve, there will be an overall double minimum potential of free energy function and the roughness of the energy landscape simply acts to modulate the rate of passing over the transition state. This transition state is actually a set of many configurations and could be said to consist of numerous microtransition states in a funnel toward the folded state. The kinetics in this situation are simple exponential. If the driving forces for folding are considerably smaller, the folding temperature can become close to the glass transition temperature. In this case one encounters considerable slowing of the folding process itself; a Type II scenario emerges in which individual pathways for folding can be dissected. Here there will be multiple exponential processes typically. The great irony, of course, is that in the situation where we can find individual pathways, folding will be typically very slow. Indeed, nearly kinetically unfoldable proteins would exhibit the most clearly defined pathway for folding. These discrete pathways, however, are not self-averaging aspects of the dynamics and are sensitive to individual mutations in sequence.

For very large driving forces, one can encounter Type 0 scenario folding in which essentially all of the dynamics goes on in a downhill manner. If a Type 0 scenario can occur much above T_g, this gives rise to processes that are very fast (of the order of ordinary homopolymer collapse times).[127] On the other hand, if the glass transition intervenes, which is likely if nonspecific collapse occurs, individual pathways can still be found, and, again, they will be strongly sequence dependent and sensitive to mutations.

If the qualitative nature of the interaction energy scales is understood, detailed temperature dependences can be obtained by the energy landscape analysis. A typical feature of this analysis is that one obtains curved Arrhenius plots for folding times, much like those actually occurring in experimental situations. This curvature reflects the roughness of the energy scale of the particular protein and enters in both a thermodynamic and dynamic way. The other energy scale is related to the folding temperature itself and to the stability gap in the energy spectrum of kinetically foldable proteins. Simple linear free energy relations between the folding time and the stability gap energy scale are obtained. A most remarkable feature, however, is that there are discontinuities in these relations and in the apparent activated energies themselves as the glass transition is approached. The main difficulty in using energy landscape analysis to interpret laboratory experiments is the temperature dependence of the underlying thermodynamic forces. Still, the self-averaging nature of the roughness energy scales versus the specific sequence dependence of the stability gap scale should allow some insight to be obtained in real experiments. The employment of different modes of denaturation will be essential in differentiating these energy scales of the protein folding landscape. One can think of the use of chemical denaturants, such as urea and guanidine, that largely bind to unfolded configurations as primarily affecting the stability gap rather than the roughness energy scale. On the other hand, pressure will strongly affect all solvent-mediated forces and thus will correlate with the roughness energy scale.[128–131]

Another complexity in laboratory experiments is

that there can be multiple order parameters for real proteins, since folded structures differ in several ways from the typical unfolded ones. The point is, however, that there are probably only a few such parameters and a few overall energy scales that are relevant. If the dynamic reorganization timescales for each of these order parameters are similar, the many reaction coordinate situation does not differ dramatically from the one effective coordinate picture we have discussed in detail in this paper. If the timescales for different motions differ appreciably, either through local energy barriers or glass transition temperatures that vary with these order parameters, a more complex scenario in which the folding bottleneck is largely independent of the equilibrium free energy barrier can arise. Still the few coordinate generalization of the present analysis would be applicable. Experimentally this situation would resemble Type II or Type 0B one coordinate scenarios, in that multiexponential kinetics would be prevalent.

The most important additional order parameters are those measuring the degree of collapse, secondary structure, e.g., helical content, and side chain ordering. The glass transition characteristics depend greatly on collapse, so this is one possible source of decoupling of the bottleneck from the equilibrium free energy barrier.[6] The ruggedness of the energy landscape also can depend on side chain orientation since some misassociations may simply not be sterically allowed for some side chain orientations. In addition, the configurational entropy of the backbone depends on its helical content, again affecting the dynamic glass transition. Certainly in multidomain proteins one must use different reaction coordinates for each folding unit. Even single domain proteins may have different folding substructures. Some analyses such as that of Bryngelson and Wolynes, suggest that the critical nucleus for folding is large,[6] but other studies suggest smaller sizes for the critical nucleus and concomitantly smaller folding units with separate reaction coordinates.[27] In any case, an energy landscape analysis allows us to reduce, in many circumstances, a huge number of variables down to only a few degrees of freedom and a statistical characterization of the roughness of the energy landscape. The true diversity of the energy landscape only comes through in the Type II scenarios in which the glass transition has intervened. A study of most experiments suggests that many proteins are near the glass transition and may show Type 0B and Type II scenarios. Since the roughness of the energy scale is self-averaging, it will be interesting to explore the phase diagrams for different proteins and especially to examine different protein compositional classes to see if there are systematic differences in energy scale roughness in in vitro folding.

One of the major fruits of the energy landscape analysis of protein folding has been a simple variational criterion for achieving fast-folding proteins. The minimal frustration principle, which at first seemed a qualitative concept, has been formulated now as a criterion for the maximization of the folding temperature compared to the glass transition temperature. This principle has already been used to reverse engineer proteins to discover correlations that are important in predicting protein structure.[83,84] In addition, it has been used to design proteins that can fold on reasonable timescales on computers.[132] It will be interesting to see whether the combination of the reverse engineering and engineering approaches will allow the design of kinetically foldable proteins in the laboratory.

ACKNOWLEDGMENTS

We gratefully acknowledge useful discussions with and encouragement from William A. Eaton, Hans Frauenfelder, V. Adrian Parsegian, and Attila Szabo. The useful comments on an earlier version of the manuscript by Hue Sun Chan, Ken A. Dill, and Donald M. Engelman are also appreciated. J.D.B. thanks the Division of Computer Research and Technology of the National Institutes of Health for supporting his research. J.N.O. is a Beckman Young Investigator. The work in San Diego has been funded by the Arnold and Mabel Beckman Foundation and by the National Science Foundation (Grant MCB-93-16186). J.N.O. is in residence at the Instituto de Física e Química de São Carlos, Universidade de São Paulo, São Carlos, SP, Brazil during part of the summers. P.G.W.'s research is supported by NIH (grant 1 R01 GM44557). This work was started at the Institute for Theoretical Physics, University of California, Santa Barbara, during a miniprogram on Protein and Nucleic Acid Folding in January 1993, supported by the NSF (grant PHY-89-04035).

REFERENCES

1. Anfinsen, C.B., Haber, E., Sela, M., White, F.H. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc. Natl. Acad. Sci. U.S.A. 47:1309–1314, 1961.
2. Levinthal, C. How to fold graciously. In: "Mossbauer Spectroscopy in Biological Systems." Proceedings of a meeting held at Allerton house, Monticello, Illinois. DeBrunner, P., Tsibris, J., Munck, E. (eds.). Urbana, IL: University of Illinois Press, 1969: 22–24.
3. Harrison, S.C., Durbin, R. Is there a single pathway for the folding of a polypeptide chain? Proc. Natl. Acad. Sci. U.S.A. 82:4028–4030, 1985.
4. Bryngelson, J.D., Wolynes, P.G. Spin glasses and the statistical mechanics of protein folding. Proc. Natl. Acad. Sci. U.S.A. 84:7524–7528, 1987.
5. Bryngelson, J.D., Wolynes, P.G. Intermediates and barrier crossing in a random energy model (with applications to protein folding). J. Phys. Chem. 93:6902–6915, 1989.
6. Bryngelson, J.D., Wolynes, P.G. A simple statistical field theory of heteropolymer collapse with application to protein folding. Biopolymers 30:177–188, 1990.
7. Bairoch, A., Boeckmann, B. The SWISS-PROT protein sequence data bank. Nucl. Acids Res. 20;2019–2022, 1992.

8. Ghélis, C., Yon, J. "Protein Folding." New York: Academic Press, 1982.
9. Pain. R.H. (ed.) "Mechanisms of Protein Folding." Oxford; Oxford University Press (IRL Press), 1994.
10. Pauling, L.C. "General Chemistry," 3rd ed. San Francisco: W.H. Freeman, 1970: 381–419.
11. Lewis, G.N., Randall, M. "Thermodynamics," 2nd ed. revised by Pitzer, K.S., Brewer, L. New York: McGraw-Hill, 1961: 138–144.
12. Franks, F. The hydrophobic interaction. In: "Water: A Comprehensive Treatise," Vol. 4. Franks, F. (ed.). New York: Plenum, 1975: 1–94.
13. Leikin, S., Rau, D.C., Parsegian, V.A. Direct measurement of forces between self-assembled proteins: Temperature-dependent exponential forces between collagen triple helices. Proc. Natl. Acad. Sci. U.S.A. 91:276–280, 1994.
14. Zwanzig, R., Szabo, A., Bagchi, B. Levinthal's paradox. Proc. Natl. Acad. Sci. U.S.A. 89:20–22, 1992.
15. Leopold, P.E., Montal, M., Onuchic, J.N. Protein folding funnels: A kinetic approach to the sequence-structure relationship. Proc. Natl. Acad. Sci. U.S.A. 89:8721–8725, 1992.
16. Dill, K.A. Folding proteins: Finding a needle in a haystack. Curr. Opinion Struct. Biol. 3:99–103, 1993.
17. McCammon, J.A., Harvey, S.C. "Dynamics of Proteins and Nucleic Acids." Cambridge: Cambridge University Press, 1987.
18. Frauenfelder, H., Parak, F., Young, R.D. Conformational substates in proteins. Annu. Rev. Biophys. Biophys. Chem. 17:451–479, 1988.
19. Poland, D., Scheraga, H.A. "Theory of Helix-Coil Transitions in Biopolymers." New York: Academic Press, 1970.
20. Landau, L.D. Lifshitz, E.M. "Statistical Physics, Part I," 3rd ed. Sykes, J.B., Kearsley, M.J. trans., Course of Theoretical Physics, Vol. 5. Oxford: Pergamon Press, 1980: 531–537.
21. Lifshitz, E.M., Pitaevskii, L.P. "Physical Kinetics." Sykes, J.B., Franklin, R.N. trans., Course of Theoretical Physics, Vol. 10. Oxford: Pergamon Press, 1981: 427–438.
22. Ma, S.-K. "Statistical Mechanics." Fung, M.K., trans. Singapore: World Scientific Publishing Company, 1985: 265–268.
23. Privalov, P.L. Stability of proteins. Small globular proteins. Adv. Protein Chem. 33:167–241, 1979.
24. Ginsburg, A., Carroll, W.R. Some specific ion effects on the conformation and thermal stability of ribonuclease. Biochemistry 4:2159–2174, 1965.
25. Anfinsen, C.B. The formation and stabilization of protein structure. Biochem. J. 128:737–749, 1972.
26. Nojima, H., Ikai, A., Oshima, T., Noda, H. Reversible thermal unfolding of thermostable phosphoglycerate kinase. Thermostability associated with zero mean enthalpy change. J. Mol. Biol. 116:429–442, 1972.
27. Thirumalai, D., Guo, Z. Nucleation mechanism for protein folding and theoretical predictions for hydrogen-exchange labelling experiments. Biopolymers, in press.
28. Binder, K., Young, A.P. Spin glasses: Experimental facts, theoretical concepts and open questions. Rev. Mod. Phys. 58:801–976, 1986.
29. Mézard, M., Parisi, G., Virasoro, M.A. "Spin Glass Theory and Beyond." Singapore: World Scientific, 1986.
30. Fischer, K.H., Hertz, J.A. "Spin Glasses." Cambridge: Cambridge University Press, 1991.
31. Anderson, P.W. The concept of frustration in spin glasses. J. Less-Common Metals 62:291–294, 1978.
32. Garey, M.R., Johnson, D.S. "Computers and Intractability: A Guide to the Theory of NP-Completeness." San Francisco: W. H. Freeman, 1979.
33. Unger, R., Moult, J. Finding the lowest free energy conformation of a protein is an NP-hard problem: Proof and implications. Bull. Math. Biol. 55:1183–1198, 1993.
34. Kirkpatrick, S., Gelatt, C.D., Jr., Vecchi, M.P. Optimization by simulated annealing. Science 220:671–680, 1983.
35. LaBean, T.H., Kauffman, S.A., Butt, T.R. Libraries of random-sequence polypeptides with high yield as carboxy-terminal fusions with ubiquitin. Preprint, 1994.
36. Davidson, A.R., Sauer, R.T. Folded proteins occur frequently in libraries of random amino acid sequences. Proc. Natl. Acad. Sci. U.S.A. 91:2146–2150, 1994.
37. Branden, C., Tooze, J. "Introduction to Protein Structure." New York: Garland Publishing, 1991: 62.
38. Presta, L.G., Rose, G.D. Helix signals in proteins. Science 240:1632–1641, 1988.
39. Richardson, J.S., Richardson, D.C. Amino acid preferences for specific locations at the ends of α helices. Science 240:1648–1652, 1988.
40. Camacho, C.J., Thirumalai, D. Kinetics and thermodynamics of folding in model proteins. Proc. Natl. Acad. Sci. U.S.A. 90:6369–6372, 1993.
41. Chan, H.S., Dill, K.A. Transition states and folding dynamics of proteins and heteropolymers. J. Chem. Phys. 100:9238–9257, 1994.
42. Chan, H.S., Dill, K.A. Energy landscapes and the collapse dynamics of homopolymers, J. Chem. Phys. 99: 2116–2127, 1993.
43. Brawer, S.A. "Relaxation in Viscous Liquids and Glasses." New York: Am. Ceram. Soc., 1983.
44. Jäckle, J. Models of the glass transition. Rep. Prog. Phys. 49:171–232, 1986.
45. Derrida, B. Random-energy model: Limit of a family of disordered models. Phys. Rev. Lett. 45:79–82, 1980.
46. Derrida, B. Random-energy model: An exactly solvable model of disordered systems. Phys. Rev. B 24:2613–2626, 1981.
47. Gross, D.J., Mézard, M. The simplest spin glass. Nucl. Phys. B 240:431–452, 1984.
48. Garel, T., Orland, H. Mean-field model for protein folding. Europhys. Lett. 6:307–310, 1988.
49. Shakhnovich, E.I., Gutin, A.M. Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach. Biophys. Chem. 34: 187–199, 1989.
50. Lau, K.F., Dill, K.A. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. Macromolecules, 22:3986–3997, 1989.
51. Lau, K.F., Dill, K.A. Theory for protein mutability and biogenesis. Proc. Natl. Acad. Sci. U.S.A. 87:638–642, 1990.
52. Chan, H.S., Dill, K.A. "Sequence space soup" of proteins and copolymers. J. Chem. Phys. 95:3775–3787, 1991.
53. Yue, K., Dill, K.A. Sequence-structure relationships in proteins and copolymers. Phys. Rev. E 48:2267–2278, 1994.
54. Shakhnovich, E.I., Gutin, A.M. Implications of thermodynamics of protein folding for evolution of primary sequences. Nature (London) 346:773–775, 1990.
55. Shakhnovich, E.I., Gutin, A.M. Enumeration of all compact conformations of copolymers with random sequence of links. J. Chem. Phys. 93:5967–5971, 1990.
56. Covell, D.G., Jernigan, R.L. Conformation of folded proteins in restricted spaces. Biochemistry 29:3287–3294, 1990.
57. Hinds, D.A., Levitt, M. A lattice model for protein structure prediction at low resolution. Proc. Natl. Acad. Sci. U.S.A. 89:2536–2540, 1992.
58. Honeycutt, J.D., Thirumalai, D. Metastability of the folded states of globular proteins Proc. Natl. Acad. Sci. U.S.A. 87:3526–3529, 1990.
59. Honeycutt, J.D., Thirumalai, D. The nature of the folded state of globular proteins. Biopolymers 32:695–709, 1992.
60. Scheraga, H.A. Influence of interatomic interactions on the structure and stability of polypeptides and proteins. Biopolymers 20:1877–1899, 1981.
61. Jones, C.M., Henry, E.R., Hu, Y., Chan, C.-K., Luck, S.D., Bhuyan, A., Roder, H., Hofrichter, J., Eaton, W.A. Fast events in protein folding initiated by nanosecond laser photolysis. Proc. Natl. Acad. Sci. U.S.A. 90:11860–11864, 1993.
62. O'Shea, E.K., Klemm, J.D., Kim, P.S., Alber, T. X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil. Science 254:539–544, 1991.
63. Gō, N. Theoretical studies of protein folding. Annu. Rev. Biophys. Bioeng. 12:183–210, 1983.
64. Schulz, G.E., Schirmer, R.H. "Principles of Protein Structure." New York: Springer-Verlag, 1979: 147–148.
65. Creighton, T.E. "Proteins." New York: W.H. Freeman, 1984: 231.
66. Flory, P.J. The configuration of real polymer chains. J. Chem. Phys. 17:303–310, 1949.

67. Flory, P.J. "Principles of Polymer Chemistry." Ithaca, NY: Cornell University Press, 1954: 523–530.
68. Ptitsyn, O.B. The molten globule state. In: "Protein Folding." Creighton, T.E. (ed.). New York: W.H. Freeman, 1992: 243–300.
69. Glasstone, S., Laidler, K.J., Eyring, H. "The Theory of Rate Processes." New York: McGraw-Hill, 1941.
70. Moore, J.W., Pearson, R.G. "Kinetics and Mechanism." New York: John Wiley 1981: 137–191.
71. Sali, A., Shakhnovich, E., Karplus, M. How does a protein fold? Nature (London) 369:248–251, 1994.
72. Ferry, J.D., Grandine, E.R., Fitzgerald, E.R. Viscoelastic relaxation of polyisobutylene. J. Appl. Phys. 24:911–921, 1953.
73. Creighton, T.E. The protein folding problem. In: "Mechanisms of Protein Folding." Pain, R.H. (ed.). Oxford: Oxford University Press (IRL Press), 1994: 1–25.
74. Rao, S.P., Carlstrom, D.E., Miller, W.G. Collapsed structure polymers: A scattergun approach to amino acid copolymers. Biochemistry 13:943–952, 1974.
75. Oxtoby, D. New perspectives on freezing and melting. Nature (London) 347:725–730, 1990.
76. Kim, P.S., Baldwin, R.L. Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding. Annu. Rev. Biochem. 51:459–489, 1982.
77. Lesk, A.M., Rose, G.D. Folding units in globular proteins. Proc. Natl. Acad. Sci. U.S.A. 78:4304–4308, 1981.
78. Karplus, M., Weaver, D.L. Protein-folding dynamics. Nature (London) 260:404–406, 1976.
79. Karplus, M., Weaver, D.L. Diffusion-collision model for protein folding. Biopolymers 18:1421–1437, 1979.
80. Weaver, D.L. Microdomain dynamics in folding proteins. Biopolymers 21:1275–1300, 1982.
81. Weaver, D.L. Alternative pathways in diffusion-collision controlled protein folding. Biopolymers 23:675–694, 1984.
82. Karplus, M., Weaver, D.L. Protein folding dynamics: The diffusion-collision model and experimental data. Protein Sci. 3:650–668, 1994.
83. Goldstein, R.A., Luthey-Schulten, Z.A., Wolynes, P.G. Optimal protein-folding codes from spin-glass theory. Proc. Natl. Acad. Sci. U.S.A. 89:4918–4922, 1992.
84. Goldstein, R.A., Luthey-Schulten, Z.A., Wolynes, P.G. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. Proc. Natl. Acad. Sci. U.S.A. 89:9029–9033, 1992.
85. Karplus, M., McCammon, J.A. Dynamics of proteins: Elements and function. Annu. Rev. Biochem. 53:263–300, 1983.
86. Šali, A., Shakhnovich, E., Karplus, M. Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. J. Mol. Biol. 235:1614–1636, 1994.
87. Rooman, M.J., Kocher, J.-P.A., Wodak, S.J. Extracting information on folding from the amino acid sequence: Accurate predictions for protein regions with preferred conformation in the absence of tertiary information. Biochemistry, 31:10226–10238, 1992.
88. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. Proteins 18:338–352, 1994.
89. Kolinski, A., Skolnick, J. Monte Carlo simulations of protein folding. II. Application to protein A, ROP and crambin. Proteins 18:353–366, 1994.
90. Socci, N.D., Onuchic, J.N. Folding kinetics of protein-like heteropolymers. J. Chem. Phys. 101:1519–1528, 1994.
91. Taketomi, H., Ueda, Y., Gō, N. Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. Int J. Pept. Protein Res. 7:445–459, 1975.
92. Gō, N., Taketomi, H. Respective roles of short- and long-range interactions in protein folding. Proc. Natl. Acad. Sci. U.S.A. 75:559–563, 1978.
93. Abe, H., Gō, N. Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins. Biopolymers 20:1013–1031, 1980.
94. Kanō, F., Gō, N. Dynamics of folding and unfolding transition in a globular protein studied by time correlation functions from computer simulations. Biopolymers 21: 565–581, 1981.
95. Miyazawa, S., Jernigan, R.L. Equilibrium folding and unfolding pathways for a model protein. Biopolymers 21: 1333–1363, 1982.
96. Kolinski, A., Skolnick, J., Yaris, R. Monte Carlo simulations on an equilibrium globular protein folding model Proc. Natl. Acad. Sci. U.S.A. 83:7267–7271, 1986.
97. Skolnick, J., Kolinski, A. Simulations of the folding of a globular protein. Science 250:1121–1125, 1990.
98. Kolinski, A., Milik, M., Skolnick, J. Static and dynamic properties of new lattice model of polypeptide chains. J. Chem. Phys. 94:3978–3985.
99. Skolnick, J., Kolinski, A. Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. J. Mol. Biol. 212:819–836, 1990.
100. Rey, A., Skolnick, J. Comparison of lattice monte carlo dynamics and brownian dynamics folding pathways of the α-helical hairpins. Chem. Phys. 158:199–219, 1990.
101. Skolnick, J., Kolinski, A. Computer simulations of globular protein folding and tertiary structure. Annu. Rev. Phys. Chem. 40:207–235, 1989.
102. Chan, H.S., Dill, K.A. Polymer principles in protein structure and stability. Annu. Rev. Biophys. Biophys. Chem. 20:447–490, 1991.
103. Guo, Z., Thirumalai, D., Honeycutt, J.D. Folding kinetics of proteins: A model study. J. Chem. Phys. 97:525–535, 1992.
104. Covell, D.G. Lattice model simulations of polypeptide chain folding. J. Mol. Biol. 235:1032–1043, 1994.
105. Shakhnovich, E., Farztdinov, G., Gutin, A.M., Karplus, M. Protein folding bottlenecks: A lattice Monte Carlo simulation. Phys. Rev. Lett. 67:1665–1668, 1991.
106. Flanagan, J.M., Kataoka, M., Fujisawa, T., Engelman, D.M., Mutations can cause large changes in the conformation of a denatured protein. Biochemistry 32:10359–10370, 1993.
107. Miller, R., Danko, C.A. Fasolka, J., Balazs, A.C., Chan, H.S., Dill, K.A. Folding kinetics of proteins and copolymers. J. Chem. Phys. 96:768–780, 1992.
108. Berry, R.S., Rice, S.A., Ross, J. "Physical Chemistry." New York: John Wiley, 1980: 1165–1167.
109. Leffler, J.E., Grunwald, E. "Rates and Equilibria of Organic Reactions." New York: John Wiley, 1963: 156–161.
110. Wells, P.R., "Linear Free Energy Relationships." New York: Academic Press, 1968.
111. Matthews, C.R. Effect of point mutations on the folding of globular proteins. Methods Enzymol. 154:498–511, 1987.
112. Matouschek, A., Fersht, A.R. Protein engineering in analysis of protein folding pathways and stability. Methods Enzymol. 202:82–112, 1987.
113. Jennings ,P.A., Saalau-Bethell, S.M., Finn, B.E., Chen, X., Matthews, C.R. Mutational analysis of protein folding mechanisms. Methods Enzymol. 202:113–126, 1991.
114. Matouschek, A., Serrano, L., Fersht, A.R. Analysis of protein folding by protein engineering. In: "Mechanisms of Protein Folding." Pain, R.H. (ed.). Oxford: Oxford University Press (IRL Press), 1994: 137–159.
115. Szabo, A. The kinetics of hemoglobin and transition state theory. Proc. Natl. Acad. Sci. U.S.A. 75:2108–2111, 1978.
116. Schellman, J.A. Solvent denaturation. Biopolymers 17: 1305–1322, 1978.
117. Eaton, W.A., Henry, E.R., Hofrichter, J. Application of linear free energy relations to protein conformational changes: The quaternary structural change of hemoglobin. Proc. Natl. Acad. Sci. U.S.A. 88:4472–4475, 1991.
118. Leffler, J.E. Parameters for the description of the transition state. Science 117:340–341, 1953.
119. Chen, B.-Z., Schellman, J.A. Low-temperature unfolding of a mutant of phage T4 lysozyme. 1. Equilibrium studies. Biochemistry 28:685–691, 1989.
120. Chen, B.-Z., Baase, W.A., Schellman, J.A. Low-temperature unfolding of a mutant of phage T4 lysozyme. 2. Kinetic investigations. Biochemistry 28:691–699, 1989.
121. Jackson, S.E., Fersht, A.R. Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. Biochemistry 30:10428–10435, 1991.
122. Jackson, S.E., Fersht, A.R. Folding of chymotrypsin inhibitor 2. 2. Influence of proline isomerization on the folding kinetics and thermodynamic characterization of the

transition state of folding. Biochemistry 30:10436–10443, 1991.
123. Nall, B.T. Proline isomerization and folding of yeast cytochrome *c*. In: "Protein Folding: Deciphering the Second Half of the Genetic Code." Washington, D.C.: AAAS, 1990: 198–207.
124. Nall, B.T. Proline isomerization as a rate-limiting step. In: "Mechanisms of Protein Folding." Pain, R.H. (ed.). Oxford: Oxford University Press (IRL Press), 1994: 137–159.
125. Radford, S.E., Dobson, C.M., Evans, P.A The folding of hen lysozyme involves partially structured intermediates and multiple pathways. Nature (London) 358:302–307, 1992.
126. Sosnick, T.R., Mayne, L., Hiller, R., Englander, S.W. The barriers in protein folding. Nature Struct. Biol. 1:149–156, 1994.
127. de Gennes, P.G. Kinetics of collapse for a flexible coil. J. Phys. Lett. 46:L-639–L-642, 1985.
128. Zipp, A., Kauzmann, W. Pressure denaturation of metmyoglobin. Biochemistry 12:4217–4228, 1973.
129. Li, T.M., Hook, J.W. III., Drickamer, H.G., Weber, G. Plurality of pressure denatured forms lysozyme and chymotrypsinogen. Biochemistry 15:5571–5580, 1974.
130. Samarasinghhe, S.D., Campbell, D.M., Jonas, A., Jonas, J. High-resolution NMR study of the pressure-induced unfolding of lysozyme. Biochemistry 31:7773–7778, 1992.
131. Royer, C.A., Hinck, A.P., Loh, S.N., Prehoda, K.E., Peng, X., Jonas, J., Markley, J.L. Effects of amino acid substitution on the pressure denaturation of staphylococcal nuclease as monitored by fluorescence and nuclear magnetic resonance spectroscopy. Biochemistry 32:5222–5232, 1993.
132. Shakhnovich, E.I., Gutin, A.M. Engineering of stable and fast-folding sequences of model proteins. Proc. Natl. Acad. Sci. U.S.A. 90:7195–7199, 1993.