Annu. Rev. Phys. Chem. 1997. 48:545–600

THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective

José Nelson Onuchic
Department of Physics, University of California at San Diego, La Jolla, California 92093-0319

Zaida Luthey-Schulten and Peter G. Wolynes
School of Chemical Sciences, University of Illinois, Urbana, Illinois 61801

KEY WORDS: folding funnel, minimal frustration, lattice simulations, heteropolymer phase diagram, structure prediction

ABSTRACT

The energy landscape theory of protein folding is a statistical description of a protein's potential surface. It assumes that folding occurs through organizing an ensemble of structures rather than through only a few uniquely defined structural intermediates. It suggests that the most realistic model of a protein is a minimally frustrated heteropolymer with a rugged funnel-like landscape biased toward the native structure. This statistical description has been developed using tools from the statistical mechanics of disordered systems, polymers, and phase transitions of finite systems. We review here its analytical background and contrast the phenomena in homopolymers, random heteropolymers, and protein-like heteropolymers that are kinetically and thermodynamically capable of folding. The connection between these statistical concepts and the results of minimalist models used in computer simulations is discussed. The review concludes with a brief discussion of how the theory helps in the interpretation of results from fast folding experiments and in the practical task of protein structure prediction.

INTRODUCTION

Ever since the birth of molecular biology it has been recognized that understanding fundamental biological processes requires insights from physical chemistry. The prediction of the global structure of nucleic acids by Watson & Crick (1), with its consequences for understanding the mechanism of heredity, and the

0066-426X/97/1001-0545$08.00

local structures of proteins by Pauling et al (2) were firmly buttressed by the then contemporary understanding of molecular forces. Conversely, the elucidation of these structures inspired considerable study by physical chemists of the thermodynamics and kinetics of the processes by which these biomolecular structures were attained in nature. Anfinsen and coworkers showed that not only the local structure but also the global three-dimensional structure of protein could be reached reliably by the protein molecule acting alone through purely physicochemical processes without any special biological machinery, using only the information in the proteins amino acid sequence (3, 4). This process seemed especially remarkable when X-ray diffraction revealed that proteins were apparently not simple repetitive structures like DNA but were compact objects with complex folds whose structure was hard to predict a priori. Understanding the way in which the one-dimensional protein sequence guides the molecule to a particular three-dimensional fold fascinated generations of molecular biologists and physical chemists, as well as physicists and mathematicians, as the protein folding problem. Progress was rapid at first, with the realization by Kauzmann of the powerful role of hydrophobic forces in folding (5) and the analysis of secondary structure formation by many workers (e.g. 6). The tremendous amount of experimentation since that time yielded a bewildering complexity of thermodynamic and kinetic results from a stunning variety of probes. In general the complex behaviors were hard to relate to any theoretical understanding of protein structure and energetics. In an attempt to organize these results scientists embraced the idea that there must be a pathway for folding (7–9), whose details must be ferreted out in all their diversity. An alternative viewpoint began to emerge in the late 1980s (10, 11). This viewpoint holds that what is most important for understanding the folding process is a global overview of the protein's energy surface. This global view will be most helpful if folding occurs through organizing an ensemble of structures rather than through only a few uniquely defined structural intermediates. If this is the case, a statistical description of the protein energy landscape can be used. Such a description can be built with tools from the statistical mechanics of disordered systems, polymers, and phase transitions in finite systems. This article reviews the progress toward a physical chemical understanding of protein folding achieved by taking this new viewpoint. Other reviews aimed at audiences of physicists (12–14) or physically oriented biochemists (15–21) cover similar or related topics and should be read for their different emphasis.

The main idea that emerges from the statistical energy landscape theory is that globally the folding landscape resembles a funnel but is to some extent rugged, i.e. riddled with traps in which the protein can transiently reside (15, 22, 23). In the early stages of folding the funnel guides the protein through many different sequences of traps toward the low-energy folded (native) structure. Here there

is no pathway but a multiplicity of folding routes. For small proteins, discrete pathways emerge only late in the folding process when much of the protein has already achieved a correct configuration. The late discrete pathway arises from trapping and in a sense reflects the possible pathology of protein folding. Much like the phenomenon of hemoglobin sickling, these late events are quite sensitive to details of protein structure and sequence, explaining a large part of the bewildering complexity seen by the early experimentalists. The simple parts of the folding process, where most of the real molecular organization is going on, occur in the early events of folding and can be described using a few parameters statistically characterizing the protein folding funnel. Until recently these events were largely unstudied in the laboratory. Fast folding is starting to be studied using NMR dynamic spectroscopy (24–34), protein engineering (35, 36), laser initiated folding (37–42), and ultrafast mixing experiments (43). A few tantalizing hints from these experiments confirming the basic validity of the funnel and statistical energy landscape notions are also reviewed here, but we cannot do justice to the experimental field in this article focusing on theory.

The organization of this review is as follows. We first review the analytical background of the statistical energy landscape theory, discussing also the relevant polymer physics. We contrast the phenomena in homopolymers, random heteropolymers, and protein-like heteropolymers that are kinetically and thermodynamically capable of folding. We then discuss the connection of these ideas with the results of minimalist models used in computer simulations.

We touch upon the connection of these results with real proteins and the interpretation of kinetic folding experiments, but we also discuss lessons from the theory that help in the practical tasks of protein structure prediction and de novo protein design. In fact, development of practical prediction technology and fundamental theories of folding dynamics have been mutually supportive when the landscape perspective is used.

THE PROTEIN FOLDING ENERGY LANDSCAPE

The energy landscape is described using statistical ideas. Are proteins random objects? Sequences of the protein adenylate kinase from the archaeal genus *Methanococcus* and the pig are shown in Figure 1, and even though the protein has essentially the same structure and function in all species, the sequence identity of the archaeal bacterium protein compared to the protein in the pig is relatively low. Sections of the archaeal proteins are only 20% identical with the pig, sequentially. Two English texts with only 20% identity would appear totally unrelated strictly orthographically but could have the same meaning. Likewise, proteins with structural homologs having such low sequence identities are not uncommon. In addition, while their sequences may appear random,

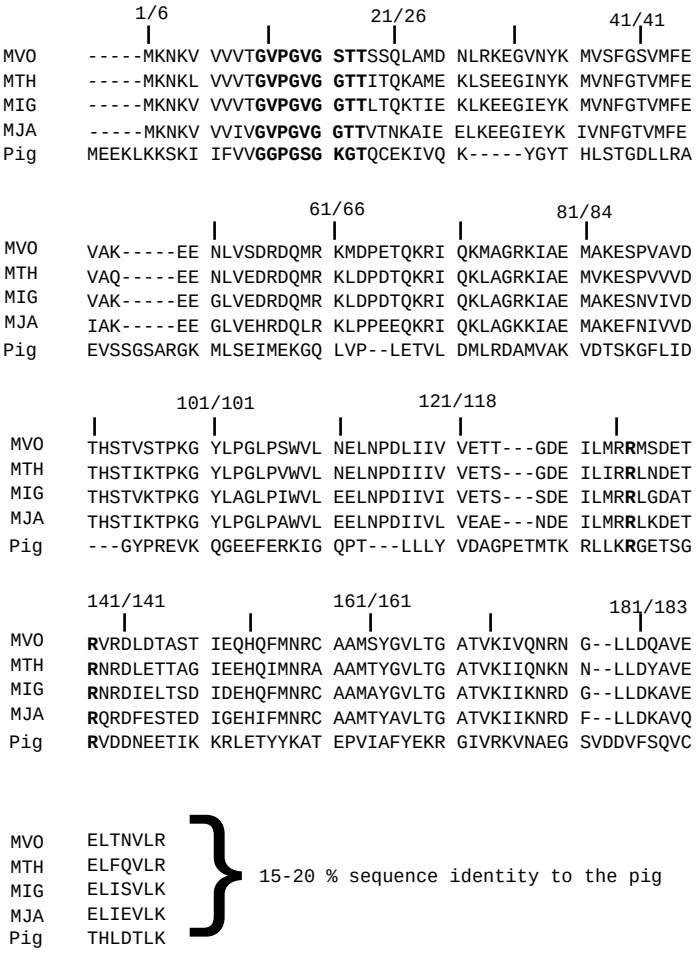

Figure 1 Block alignment of highly related archaeal adenylate kinases (AKs) from the genus *Methanococcus* to the sequence of the pig. Although the archaeal adenylate kinases, the mesophile *M. voltae* [MVO], moderate thermophile *M. thermolithotrophicus* [MTH], and extreme thermophiles *M. jannaschii* [MJA] and *M. igneus* [MIG] have 68–81% sequence identity with each other, they have only low levels of sequence identity (15–20%) to the eukaryotic and eubacterial AKs and lack several active site residues thought to be essential. The alignment is based on the energy function by Koretke and coworkers (229, 231) that was optimized using energy landscape analysis, and a model structure based on this alignment is shown in Figure 17. The highlighted residues are in the active site and play an important role in the enzymes's function to biosynthesize ADP from ATP and AMP.

proteins in nature contain many symmetrical structural elements. The difficulty in extracting the meaning from protein sequences is in discerning what features are common to all sequences, what features are specific to protein-like sequences, and what features are specific to a given structure. A view of how the ensemble of protein-like sequences is embedded in the ensemble of random heteropolymers is sketched in Figure 2. Within the space of random heteropolymers based on the 20 naturally occurring amino acids, we need to differentiate

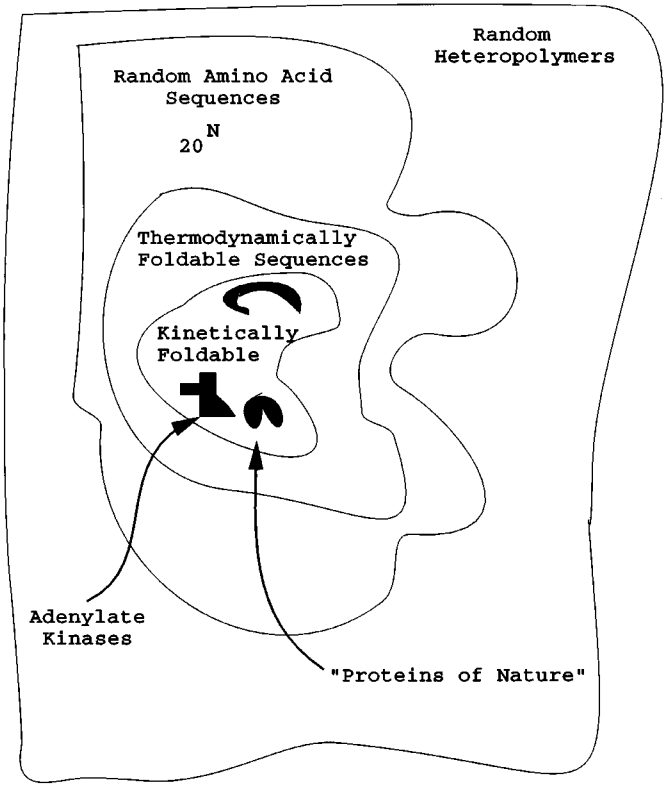

Figure 2 How proteins in nature are embedded in the ensemble of random heteropolymers. Within the space of random heteropolymers based on the 20 naturally occurring amino acids there would be 20^N possible sequences of length N. Only a small subset of these sequences are thermodynamically and kinetically foldable on the appropriate biological timescale to be seen in nature. We show the kinetically allowed set as lying within the thermodynamic one. Depending on the temperature (here taken to be rather low), it may be useful to consider the situation as reversed, i.e. a random heteropolymer can transiently touch down in the native structure but not remain there at higher temperatures.

further between thermodynamically foldable sequences and the subset of kinetically foldable sequences that make up proteins in nature. Families such as the adenylate kinase sequences belong in this last category. Determining the precise evolutionary constraints on kinetics is subtle. Organisms live at widely different temperatures (the pig around 300 K and the methanococcus at 350 K, a big difference when the properties of the solvent water are taken into account). Also the kinetic constraints especially may depend on the chain length—most large natural proteins have a domain structure. Thus the border between thermodynamic and kinetic foldability subset of sequences is a bit fuzzy. Nevertheless proteins in nature, while only marginally stable (5–12 kcal) and easy to denature with either heat or pH, must fold on a time scale that is relevant for the biological processes occurring in cells. This time is relatively short—less than a minute. This seems paradoxical at first given the many conformations that a protein can theoretically be in during folding. Quantitatively the kinetics relevant for addressing this issue can be explored by studying the formation of local structure, such as helix formation and the mechanism of collapse. But most important is understanding the funneled nature of the landscape for topological rearrangements that determine the uniqueness of the global fold (23).

Energy Landscape of a Random Heteropolymer

Many aspects of the folding process can be understood from studying the energetic properties of a random heteropolymer (RHP). Lattice models of protein conformations (see below) have been helpful in these investigations, and the simplest of these makes use of two kinds of residues (i.e. hydrophobic and hydrophilic amino acids) randomly distributed. This is a useful model for visual illustration, although it has some special properties that make it different from the more general 20–amino acid case. From both theoretical calculations and simulations we know two basic facts about the random heteropolymer: A modest structural change gives rise to a large change in energy, and low energy states that are very different in structure but close in energy exist.

In the general case, the energy of any compact conformation of such a RHP is a sum of random interactions that give rise to a rough energy landscape like the Alps. Since the energy contributions can either be stabilizing or destabilizing, the RHP is a frustrated system. In 1987 Bryngelson & Wolynes (10) applied the random energy model (REM), developed by Derrida (44, 45) to describe spin glass systems, to folding proteins (biopolymers), in particular to the misfolded states of a protein. The basic validity of this approach as a starting point has since been borne out by numerous analytical and numerical studies making the REM a zeroth order approximation for understanding random biopolymers. A brief review of the basic features of the REM for the RHP is shown in Figure 3. As a result of the random interactions the density of states is approximately a

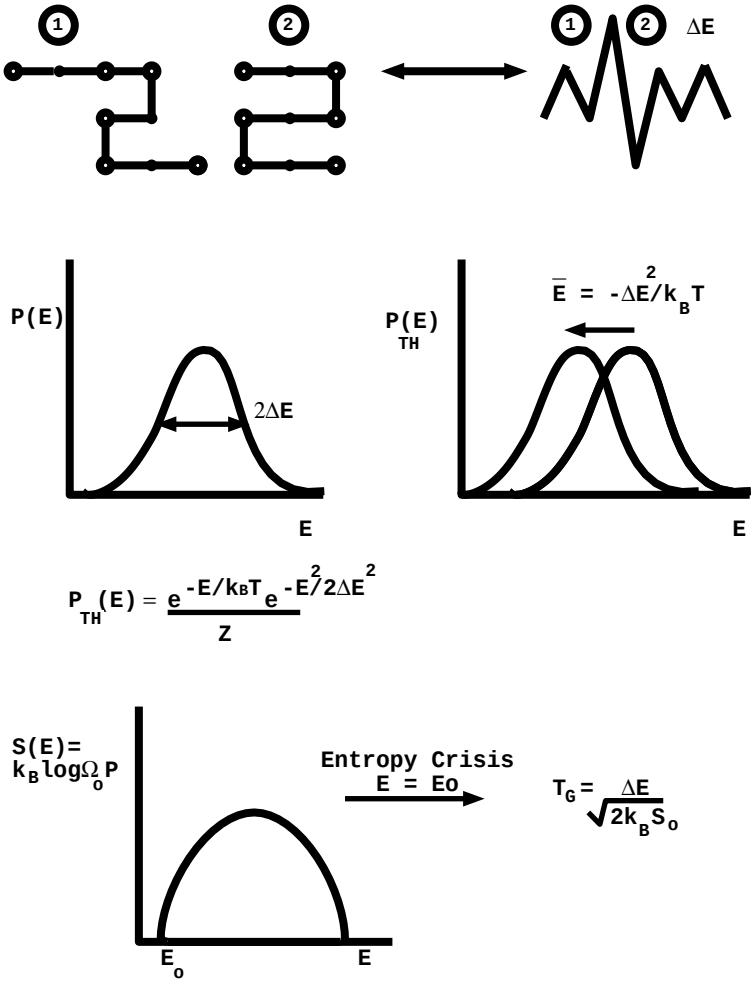

Figure 3 The energy landscape of a random heteropolymer. (*Top*) The low-energy structures of a RHP, represented by a lattice with two kinds of residues, are unrelated and the conformational changes are associated with a fluctuation $\sqrt{\Delta E^2}$ in the energy. (*Middle*) As a result of the random interactions, the density of states and the thermally weighted probabilities are approximately Gaussian with the latter centered about the mean $E = -\Delta E^2/2k_BT$ and can be modeled by random energy approximation. (*Bottom*) The system runs out of entropy when the average energy falls below E_o, and this entropy crisis is characterized by a glass transition temperature T_G, which depends on the corresponding conformational entropy and fluctuations.

Gaussian distribution with a variance ΔE. The thermally weighted probability is again a Gaussian distribution centered about the mean $E = -\Delta E^2/2k_B T$. The density of states cannot really be Gaussian in its tail, but runs out of low energy states. The entropy S is defined

$$S(E) = k_B \log [\Omega_o P(E)], \quad 1.$$

where Ω_o is the number of conformational states of the polymer. As the system is cooled the energy falls. The system runs out of entropy when the average energy falls below a critical value $E \leq E_o$ such that $S(E_o) = 0$. This entropy crisis occurs at a glass-transition temperature T_G where

$$T_G^{-1} = \sqrt{2k_B S_o/\Delta E^2} \quad 2.$$

and $S_o = k_B \log \Omega_o$. Below T_G the kinetics of the system exhibits glassy-like behavior that depends on the history of the system. Above T_G the system behaves like a viscous liquid. The slowing of transition rates between different low-energy states upon cooling leads to a strong generally non-Arrhenius temperature dependence of the rate of exploring configuration space. A more detailed discussion of the kinetics on a REM landscape is given below.

Simplest Viable Protein Folding Landscape

The folding landscape of proteins is necessarily rugged because biomolecular chains can sample many conformations during their motions and have the possibility of making inappropriate contacts between residues. In the simplest model one can assume that when nonnative contacts are made the energy contributions are random, and these contributions to the protein's energy can be treated like those for a RHP. In the ensemble of misfolded collapsed states with little native structure, the energetics can be described crudely by the REM shown in Figure 3. Low-energy structures will appear unrelated, and conformational changes are associated with a fluctuation $\sqrt{\Delta E^2}$ in the energy.

Because native contacts and local conformation energies are more stabilizing than expected, there is a smooth overall slope of the energy landscape toward the native structure. This more realistic model considers the protein to be a minimally frustrated heteropolymer. This means that the rugged landscape of real protein folding is not globally flat with totally unpredictable fluctuations about the mean as it would be for a RHP, but has a preferred direction of flow. It can be described as a rugged funnel, shown in Figure 4, whose shape can be estimated using theory and experiment. At the bottom of the funnel there is a topologically nearly unique native state. As has been emphasized by Frauenfelder and coworkers (46, 47), the physiologically active state is not just this lowest energy one but actually a whole tribe of states that differ at least in side-chain orientations (48) but possibly also in occasional topological defects

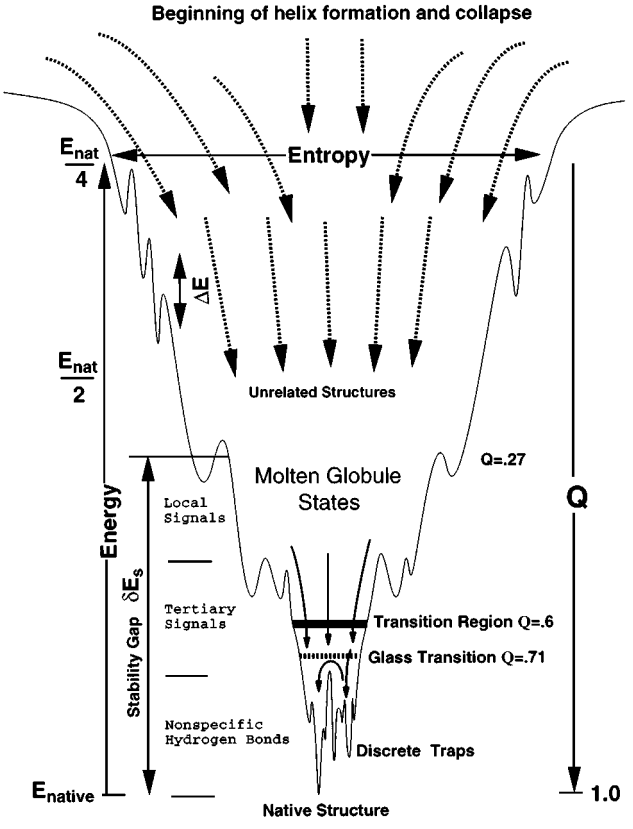

Figure 4 A viable protein folding landscape. The rugged folding landscape of a small helical protein is funnel-like, with a preferred direction of flow toward a unique native state (23, 60). The ensemble of conformations in the upper part of the funnel can be described by modified theories of a RHP (shown in Figure 3) that take into account the formation of secondary structure. The order parameters E, the solvent-averaged energy, and Q, the fraction of native-like contacts, describe the position of an ensemble of states within the funnel and stratify the landscape. The fluctuations ΔE and the stability gap δE_s between the compact misfolded or molten globule states and the native state are functions of these order parameters and can be estimated using theory and experiment. This figure is drawn after that in Onuchic et al (23) and also indicated are the contributions of local signals and tertiary contacts as well as hydrogen bonding to the stability gap, which provides the specificity of folding according to the estimates by Saven & Wolynes (74).

from this idealized structure, which is the target of so many purely structural studies. Useful order parameters to describe the position of an ensemble of states in the funnel are the solvent-averaged energy E, the fraction of native-like contacts Q, the percent of correct dihedral angles in the protein backbone A, and the percent correct secondary structure. Other order parameters such as total and local helicity may also be used to classify ensemble of states and are accessible to experimental measurement.

Through these order parameters, the folding landscape is stratified. Within each stratum we can define an average energy $\bar{E}(Q)$, although there are still many states with different energies. To describe their distribution and properties we apply a REM model. The late stages of protein folding will have few states all highly similar to the native. These could be given specific names (if necessary), and the detailed kinetics of transitions between them counts. These are analogous to the taxonomic substates discussed by Frauenfelder et al (49). Again there are further levels of conformational substates below this point, but they are usually sampled quickly at room temperature. As indicated in Figure 5, some routes can dead-end in low-energy misfolded conformations from which the protein has to partially unfold to reach the native state. In the early stages of folding, corresponding to a nearly denatured protein with $Q \approx 0$, there will be many states, and the ensemble language is clearly most appropriate. The hopping rate between microstates at each stage will depend on the ruggedness.

The complete statistical mechanical treatment of folding requires knowledge of all thermodynamic variables as a function of the order parameters, in particular the functional dependence of the thermal average energy $\bar{E}(Q)$, the ruggedness $\sqrt{\Delta E^2(Q)}$, the density of states $\Omega(E, Q)$ or equivalently the entropy S(E,Q), and the local glass transition temperature $T_G(Q)$. Using the simplest form of the random energy approximation, Bryngelson & Wolynes (10) derived these quantities for the case in which correlations of the energy states within a stratum are neglected. The energy of a given misfolded state arises from the contributions of many random terms, so the probability distribution of energies at any stratum of the funnel is a Gaussian centered about the mean energy

$$P(Q, E) = \frac{1}{\sqrt{2\pi\,\Delta E^2(Q)}} \exp\left\{-\frac{[E - \bar{E}(Q)]^2}{2\Delta E^2(Q)}\right\}. \qquad 3.$$

In a REM the probability that any two states have energy E_1 and E_2 is the product $P(E_1)P(E_2)$. If there are γ configurations per residue for a protein in its unfolded state, then the total number of configurations for a protein with N residues is $\Omega_o = \gamma^N$. In models that use a reduced description of the protein and include only the backbone coordinates, γ is less than 5, and when corrections are made for the excluded volume effect in compact configurations, $\gamma = \gamma^* \approx$

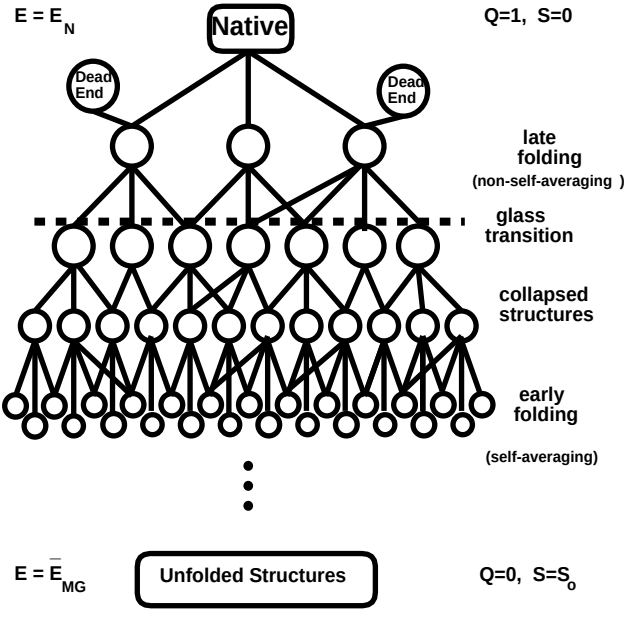

Figure 5 A schematic representation of the ensemble of folding pathways toward the native state. The native structure is in the top of the figure, and clearly the number of available conformations reduces as Q increases. Before the folding protein reaches the glass transition, there are many accessible paths between conformations. In this regime, different folding events go through different paths toward the native state. After the glass transition is reached, only a few paths remain accessible. Thus the connectivity between configurations is reduced, and several paths may lead to dead ends instead of to the native conformation.

1.5 (15, 50). As the structures become more similar to the native protein, the total number of configurations will decrease, since only a single backbone conformation represents the native state. If we call $\Omega_o(Q)$ the number of structures with similarity measure Q to the native structure and $S_o(Q)$ the corresponding entropy, then the density of conformational states with energy E and similarity Q is then

$$\Omega(E, Q) = \Omega_o(Q)P(Q, E), \qquad 4.$$

and the total entropy is

$$S(E, Q) = S_o(Q) - k_B \frac{[E - \bar{E}(Q)]^2}{2\Delta E^2(Q)}. \qquad 5.$$

At thermal equilibrium, the most probable energy is just the maximum of the

Boltzmann weighted distribution in Figure 3

$$E_{\text{m.p.}}(Q) = \bar{E}(Q) - \frac{\Delta E^2(Q)}{k_B T}. \quad 6.$$

The number of thermally occupied states and entropy associated with this most probable energy are

$$\Omega(E_{\text{m.p.}}, Q) = \exp\left[\frac{S_o(Q)}{k_B} - \frac{\Delta E^2(Q)}{2(k_B T)^2}\right] \quad 7.$$

$$S(E_{\text{m.p.}}, Q) = S_o(Q) - \frac{\Delta E^2(Q)}{2k_B T^2}. \quad 8.$$

By combining Equations 6 and 8, the free energy of the misfolded structures with configurational similarity Q and at a fixed temperature becomes

$$\begin{aligned} F(Q, T) &= E_{\text{m.p.}}(Q) - TS(E_{\text{m.p.}}, Q) \\ &= \bar{E}(Q) - \frac{\Delta E^2(Q)}{2k_B T} - TS_o(Q). \end{aligned} \quad 9.$$

Under some thermodynamic conditions, folding can be considered a two-state reaction (see Figure 6). In this case the free energy has a double minimum, with one minimum near the folded state $Q \approx 1$ and the other at the position $Q_{\min} \approx 0$ corresponding to an ensemble of collapsed misfolded states with varying degrees of ordering. The precise origin of the barrier is a subtle point involving at least the polymer physics of entropy loss on contact formation in lattice models but perhaps also explicitly cooperative many-body forces for real proteins. To a first approximation, we can neglect the entropy of the folded state so that its free energy is equal to its internal energy, E_N. At the folding temperature, T_F, the probability of being in the folded state is equal to the probability of being in the misfolded state implying, $F_{\text{native}} = F(Q_{\min}, T_F)$. This equality yields an expression for the slope of the funnel

$$\delta E_s / T_F = S_o + \Delta E^2(Q_{\min})/2k_B T_f^2 \quad 10.$$

in terms of the stability gap $\delta E_s = \bar{E}(Q_{\min}) - E_N$. Since $Q_{\min}$ is close to the unfolded state, we consider the folding temperature as being referenced to a set of states with little structural similarity to the native state, $Q \approx 0.2$–0.3 for the simple funnel in Figure 4.

Recall that a glass transition occurs at the temperature where there are too few states available, so the system remains frozen in one of a few distinct states. Within each stratum this is characterized by an entropy crisis where $S(T_G, Q) = 0$. Using Equation 8, the local glass transition temperature is

$$T_G(Q) = \sqrt{\frac{\Delta E^2(Q)}{2k_B S_o(Q)}}. \quad 11.$$

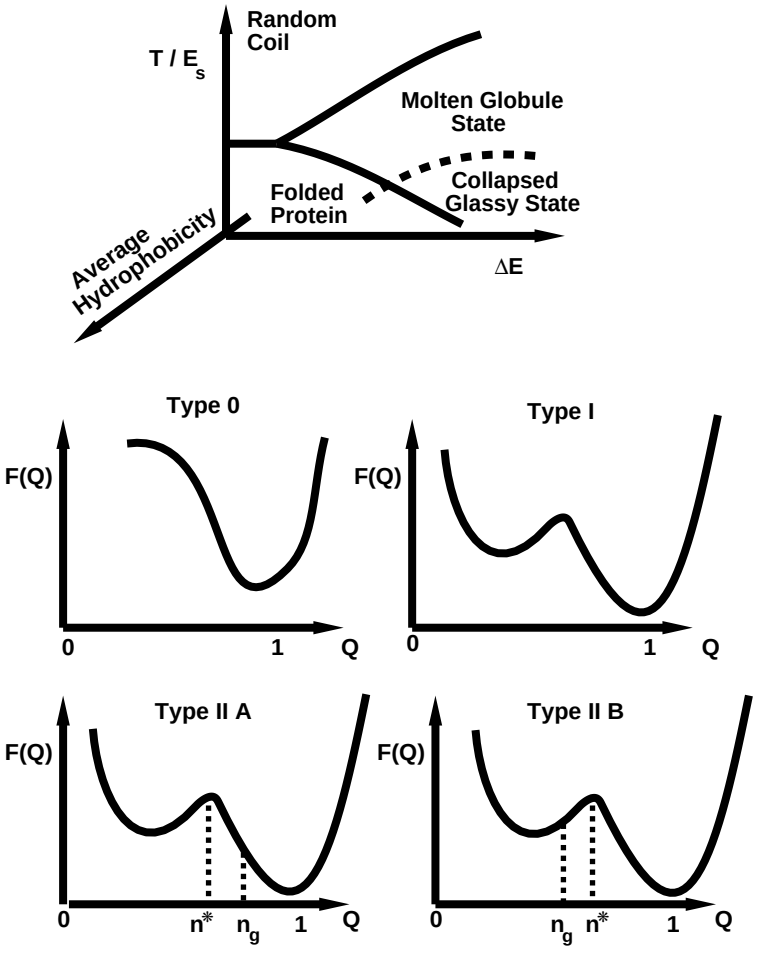

Figure 6 Phase diagram and folding scenarios according to the minimally frustrated REM analysis. (*Top*) The phase diagram along a line of some average sequence hydrophobicity shows the possible thermodynamic states of a protein modeled as a minimally frustrated heteropolymer. Varying the hydrophobicity by changes in solvent and temperature can modify the number of phases observed in the folding process. (*Bottom*) The free-energy curves consistent with the above phase diagram exhibit various folding scenarios: Type 0 and I are examples of fast folding, downhill with no barrier and a small barrier in the latter. These are typical scenarios for well-designed proteins at temperatures below the folding temperatures and conditions such that the glass transition is not encountered. The Type II scenarios occur at the right-hand side of the phase diagram under conditions that favor the formation of the glassy state either after or before the thermodynamic transition barrier.

Local glass transition temperatures are manifested in the folding and collapse times measured in lattice calculations that are summarized below. Analytical and numerical studies on lattice models have shown that the ratio of T_F/T_G can be used to distinguish fast and slow folding sequences. This ratio also plays a central role in developing energy functions to predict protein structures. Calculating this ratio using the set of states with the least structural similarity to the folded state gives (51)

$$\frac{T_F}{T_G} \approx \frac{\delta E_s}{\Delta E} \cdot \sqrt{\frac{2k_B}{S_o}}. \qquad 12.$$

As we show below, for a protein to fold, T_F/T_G must be greater than 1 for fast folding, and since S_o, ΔE^2, and E_N all depend linearly on the chain length N, T_F/T_G is independent of length and sensitive to the interaction energies.

A phase diagram is a useful tool for summarizing which states of a protein are involved in the various folding scenarios. So far our analysis has only used a single parameter Q to characterize the changes in free energy and the differences between the native and unfolded states. Clearly there are other parameters besides the number of correct contacts that could be used to compare structures and describe the partial ordering that occurs as the protein folds. For example, in a folded protein, the core consists primarily of hydrophobic residues and the surface of hydrophilic residues (52). This ordering is due to the hydrophobic effect arising from folding a protein in water. Variations in the solvent properties will have profound effects on the interaction energies of the hydrophobic groups. The phase diagram in Figure 6 shows the possible thermodynamic states of a model protein as a function of temperature and the roughness of the energy landscape expected from the minimally frustrated REM analysis (15). The phase diagram is actually a slice through a more complicated diagram along a line of some average hydrophobicity of the sequence. Since average hydrophobicity itself depends on solvent and temperature, under some conditions the coexistence curve between the random coil and the folded state disappears, and the folded state becomes only accessible after nonspecific collapse. The thermodynamic dependence of hydrophobic forces is one of the main complicating features in relating the theoretical phase diagrams (that assume temperature-independent forces) to experiment. This is most manifest in the phenomena of cold- and pressure-induced denaturation. The glass transition, which occurs after the collapse of the system, is a continuous transition. This portion of the phase diagram is typical of RHPs (53). Recent lattice simulations of Socci & Onuchic (54) on protein-like sequences probe the phase diagram as a function of temperature and average hydrophobicity and provide qualitatively the same picture.

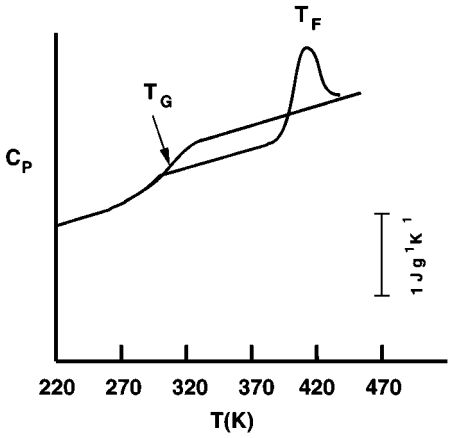

Figure 7 The heat capacity C_P of a dense hydrated protein (legumin) before and after denaturation. The heat capacity curves are based on the data from Sochava & Smirnova (58) and discussed by Angell (59). The experimental curve upon heating the native protein exhibits a peak at the folding temperature T_F around 410 K, while the bulk denatured protein cannot refold but undergoes a glass transition at $T_{G,bulk} = 320$ K, a much lower temperature. The ratio of $T_F/T_G = 1.3$ is probably an underestimate for the value of individual proteins that are highly solvated.

The phase diagrams of real proteins in the laboratory can be more complex and in fact are difficult to determine completely. The glass transition in the molten globule (MG) phase of globular proteins, which is most important for theoretical considerations in particular, has hardly been probed. The glass transition within folded proteins has been studied extensively (55, 56), but questions have been raised concerning the role of solvent dynamics in that case (57). On the other hand, the glass transitions in bulk protein materials have been directly observed (58) (see Figure 7). This has been carefully discussed by Angell recently (59). The experimental heat capacity of a native protein shows a peak when it denatures at the folding temperature. When bulk denatured protein is cooled, on the other hand, the aggregated protein undergoes a glass transition at a lower temperature, $T_{G,bulk}$. Theory would suggest that this bulk glass transition is higher than the corresponding transition for a globule. Thus the experimental data strongly support the minimal frustration principle notion that T_F exceeds T_G.

This phase diagram can be used to understand the free-energy behavior in the various folding scenarios. In general, in each region of the phase diagram, the free energy is either unimodal or bimodal. The unimodal case is called Type 0. In a Type 0A scenario no glass transition is encountered, and the folding process can be considered strictly downhill. In a Type 0B scenario the glass

transition is encountered before complete folding, and escape from individual traps determines the rate. Type I scenarios correspond with the bimodal free energies diagrams. Type 0A and Type I scenarios dominate the left-hand part of the phase diagram, in which no glass transition occurs, and the system is at a temperature such that the global minimum is the native folded state with $Q = 1$. Direct folding from the random coil state favors these scenarios, since glassy states can only occur in nearly collapsed chains. Type II scenarios occur in the right-hand part near the coexistence curves for the folded, collapsed, and collapsed frozen states. In a Type IIA scenario the glass transition occurs after the thermodynamic barrier, and in a Type IIB scenario the folding protein becomes glassy before the barrier is reached.

MICROSCOPIC APPROACHES TO THE ENERGY LANDSCAPES OF PROTEINS AND HETEROPOLYMERS

The energy landscape view of protein folding addresses a global characterization of the states involved in folding through statistical means. A central problem for physical chemists is to see how these global characteristics are related to the microscopic forces. This is difficult because proteins are chemically complex. Their backbones contain both specific and generic stereochemical information, and the side chains themselves are not particularly simple from an organic chemical viewpoint. Atomistic simulations provide one route to accommodate this complexity. Computational difficulties limit the extensive use of this approach at present. Thus most exploration of the energy landscape ideas has used very simple so-called minimalist models of the interactions to get started. The connection of microscopic forces to the landscape have been carried out in two divergent styles: One important thread uses the formal statistical mechanics of disordered systems to proceed. The advantage of this approach is that it enriches the conceptual tools of the energy landscape framework for interpreting experiments and provides a set of algorithms that can be quantitatively used. The other thread of approach investigates lattice models and off-lattice models (see below) of heteropolymers by using computer simulation. The advantage of this approach is that it provides controlled tests of the theory, by accessing a broad range of regimes not necessarily found in a laboratory, and perhaps most significantly visual representations that make the ideas concrete for everyone. In this section we review the formal statistical mechanical approaches and some of the new ideas that come into play in even simple microscopic theories.

The REM can itself be used to relate landscape characteristics to the microscopic forces. At first, the assumed lack of correlation between configurations would seem to be a severe approximation. To be at all meaningful, characteristics of the complete energy landscape of a heteropolymer must be

described with additional order parameters that quantify the way in which the energy landscape is stratified (15, 60). This is quite different from the situation where the REM is applied to magnetic spin glasses (44). The primary quantitative parameters of the energy landscape model, the entropy and the fluctuations in energy, depend on several gross characteristics of partially folded proteins. The most important additional order parameter is the degree of collapse of the heteropolymer. Much of the work by Dill and coworkers has focused on this topic (61). In the energy landscape context we noted that frustration arises from nonlocal interactions. Local interactions can be satisfied individually by local adjustments in a strictly one-dimensional model. Thus a fully extended chain does not have a rugged energy landscape. Bryngelson & Wolynes showed how the ruggedness depends on the overall density of a polymer globule (62). Using a simple Flory theory of the entropy of such a globule as a function of its degree of collapse, they showed that the glass transition is strongly coupled to the collapse transition and, indeed, can only occur once collapse is thermodynamically favorable. Lattice simulations by Socci & Onuchic support this notion (54). An important qualitative result of their analysis was that the search difficulties that can only occur because of landscape ruggedness are also already partially simplified by the reduction in entropy due to collapse. Therefore collapse has somewhat contradictory effects on rates of folding and conformational search. The observation that the glass transition occurs only in collapsed states also simplifies the approximate treatment of glass transitions for interaction potentials with a realistic level of complexity. In these engineering applications of energy landscape theory, discussed in the previous section, highly collapsed structures can be generated on the computer or modeled as alternate native protein structures from a database. This has allowed the use of the REM approximation to crudely locate transition temperatures in such models and to design optimal energy functions for structure prediction, as we discuss below.

Stratified random energy models can also be used to take into account other order parameters in folding. At least partially because of its name, secondary structure has always played an important role in the discussion of folding. The helix-coil transition, as studied by the classical theories of the 1950s (6), can play an important role even before the protein molecule collapses. Peptide fragments often have a weak tendency to form their final correctly folded structure autonomously or to form helices even when ultimately they will not in the folded protein (63, 64). Qualitative notions such as the framework model and the diffusion-collision hypothesis advanced the idea that secondary structure formation, nearly on its own, could account for the physical basis of folding. The modern analysis shows this is not sufficient, but secondary structure formation can be ignored only at great peril (23, 65). The molten globule state of many proteins contains considerable amounts of local secondary structure,

such as α-helices. This is sometimes the case even when the final protein does not contain helices (66). Desolvation makes this necessary in order to satisfy the strong constraints of hydrogen bonding. The energy landscape of a polymer therefore depends on the amount of helix present as well as on its degree of collapse. It is possible to simultaneously take into account helix-coil transitions and collapse in a REM for the heteropolymer (67). The local interactions of helix formation are one dimensional. The effect of hydrogen bonding on entropy can be taken into account through the use of two order parameters: one measuring the amount of helix, the other measuring the number of defects in the helical structure. The finite energy cost of a defect in a one-dimensional system is the primary cause of the rounding of the helix-coil transition relative to a true phase transition in three dimensions. In the absence of collapse the combinatorics associated with these order parameters is straightforward. New phenomena enter because of the interplay between the local polymer configuration and collapse. An entirely helical chain cannot take advantage of a general hydrophobic attraction or advantageous random contacts, so it will not be collapsed. If we know the polymer is collapsed through these interactions, we must take into account how the entropy of the chain is reduced through excluded volume and its confinement to a globular structure.

When these effects are taken into account for rigid chains, it becomes clear that the polymer will also order in a fashion expected for liquid crystals forming bundles of helices most favorably. These effects can be taken into account using theories of polymeric liquid crystal transitions by Flory and coworkers (68–70), Onsager (71) and Grover & Zwanzig (72). This analysis makes clear that even for the homopolymer, the secondary structure and tertiary structure can be in conflict and exhibit frustration. This was also seen in models by Bascle et al (73) of the hydrophobic homopolymer collapse with secondary structure. The primary effect of these partial orderings is to considerably reduce the configurational entropy of the chain even before heteropolymer effects are taken into account. This is of great importance in describing the global landscape of real proteins as we discuss below.

Still more order parameters stratifying the energy landscape can be taken into account while otherwise keeping to the REM level of approximation. One of the more important of these is the fraction of secondary structure that is locally correct. Locally correct secondary structure might be induced by local conformational signals in the sequence. Prolines, for example, often break up helices while end capping sequences for helices have been proposed. Saven & Wolynes (74) have shown how such local conformational signals are more effective in the collapsed helical state because of the already large degree of entropy diminution caused by collapse with hydrogen bonding. These amplification effects arise from the cooperative loss of entropy density caused by

collapse. If minimally frustrated, such local signals should not conflict greatly with tertiary structure guiding forces. Using peptide studies to estimate the bare local signal strength predicts that one third of the energetic structural specificity can come from this source.

The microscopic approach based on the REM can be taken somewhat further through the use of a generalized version of the REM in which there are uniform correlations in the landscape. These are not describable by global order parameters because they refer to a triangulation of the landscape—each pair of structures has some degree of correlation that describes their similarity in energy. A scheme that takes into account the pair correlations of energy levels was introduced for spin glasses by Derrida and is known as the generalized REM (GREM) (75, 76). It can be used to better approximate transition temperatures but also to characterize more completely the basins of attraction once freezing occurs on the landscape.

Plotkin et al have developed the correlated landscape approach for the RHP (77). The primary new ingredient needed in the theory is a treatment of the entropy of a chain as a function of its fractional similarity to another chain configuration in terms of pair contacts. This problem resembles the statistical mechanics of vulcanization, and many useful approximations from that theory can be used and extended (50, 78, 79). The essential physical result is that the first few contacts to be organized in folding are more entropically costly than later ones. The log number of basins is reduced through correlations by 70%, thus limiting the difficulty of the search problem at the glass temperature. The theory smoothly goes over to the REM in all details when highly cooperative many-body forces are dominant, but one important quantitative result is that even for the worst case of pair interactions alone, the glass transition temperature is actually very little changed by the correlations, thus justifying the use of simpler theory for many quantitative uses such as the structure-prediction algorithm (51, 80). Also, Pande, Grosberg, and coworkers have shown that correlated landscapes are important when studying minimalist folding models (81).

The correlated landscape model can be easily extended to deal with the introduction of the minimal frustration necessary to obtain a funneled landscape. It is necessary only to postulate a priori that a particular structure for the given sequence is much more stable than the typical GREM ground state. Plotkin and collaborators have used this to find the form of the free energy as a function of the topological order parameter Q on a funneled landscape (82). An important question highlighted in this approach is the physical origin of the thermodynamic barrier to protein folding under given conditions. In the simple pair interaction model, their barrier is rather low and is largely entropic (82). Quantitative treatment of the barrier height requires consideration of the coupling to the overall degree of collapse and suggests that it is important

to consider the partially folded protein as having a core-halo structure with a high-density well-folded interior surrounded by a low density halo. The addition of many-body forces that mimic surface area models of hydrophobicity leads to larger thermodynamic (largely energetic) barriers between folded and denatured states in this treatment.

An extremely important statistical mechanical tool for investigating the connection between the microscopic forces and the energy landscape of proteins in heteropolymers has been the replica method from spin glass theory (83). The method was first used for polymers but truly flowered in the study of random magnetic systems. This technique has some forbidding mathematical aspects both for the neophyte and for the initiated. It has, however, been studied extensively in the statistical mechanics of magnetic spin glasses, and a reasonably good physical feeling for the meaning of its results can be obtained in the study of those magnetic systems. A desirable feature of the technique is that it puts the theory of protein and heteropolymer landscapes on the same footing as these other phase transition systems for which the universality classes of the ordering transitions are believed to be understood.

Secondly, while the algebra is sometimes complicated, the manipulations used in replica methods resemble in many ways the standard field theoretical techniques used in the condensed matter physics of homogeneous systems, so a variety of standard approximation methods can be used, albeit with some care. In addition, the properties that have been understood for spin glasses below their transition temperatures can be used to discuss non–self-averaging behavior of proteins and heteropolymers. Indeed, it turns out that several questions that are abstract and difficult to test experimentally for spin glass phase transitions of magnets turn out to be extremely important and particularly apropos in the experimental study of proteins. It is likely that some of the fundamental and surprising results from the theory of spin glasses below T_C, such as ultrametricity (84), will first be tested in the context of biological heteropolymers.

Early treatments used replicas only to study the role of disorder in expanded phases where the excluded volume problem was the interesting question (85–88). The first treatments of RHPs addressing folding with replica methods were done by Garel, Orland, and coworkers (12, 89, 90) and Shakhnovich & Gutin (91). They started with the partition function of a connected polymer interacting with random interactions, sometimes supplemented with three-body interactions that prevent nonphysical collapse of the polymer chain and allow the inclusion of torsional effects. The basic partition function for a single chain with a given sequence of N amino acid residues and, therefore, a single set of interactions has the simple form

$$Z = \int \prod_{i=1}^{N} d\vec{r}_i \prod_{i=1}^{N} \delta[(r_{i-1} - r_i) - a] \prod_{i<j} e^{-\beta V_{ij}(r_{ij})}, \qquad 13.$$

where $\beta = (k_B T)^{-1}$. The vectors $\vec{r}_i$ denote the position of selected backbone atoms, typically C_α atoms, and the δ function product imposes the chain connectivity constraint, e.g the distance between consecutive C_α values is $a = 3.8$ Å. For lattice simulations this term would be modified to specify the lattice spacings. The monomers interact only when they are in contact, so $V_{ij} = v_{ij}\delta(r_i - r_j)$ and the individual v_{ij} values were chosen as independent Gaussian random variables. This partition function is analyzed in the continuum limit familiar in polymer theory, thereby giving an expression for the single chain partition function as a path integral:

$$Z = \int D\vec{r}(s) \exp\left\{ -\frac{1}{2a^2} \int_0^N ds \left[\frac{d\vec{r}(s)}{ds} \right]^2 - \beta/2 \int_0^N ds \int_0^N ds' v_{ss'} \delta[r(s) - r(s')] \right\}. \quad 14.$$

It is at this point that the replica technique is introduced in order to carry out the average of the free energy that is the logarithm of Z over random sequences:

$$\langle \log Z \rangle = \lim_{n \to 0} \frac{\langle Z^n - 1 \rangle}{n}. \quad 15.$$

The strange mathematical aspect of the replica theory for $\langle \log Z \rangle$ is that it involves taking an apparent unphysical limit of a very unusual partition function. For fixed n, Z^n is the partition function of n noninteracting copies of the same polymer system. This must be computed for general values of n; then finally the limit of n going to zero is taken. Many analytic continuations are apparently possible. The convenient feature of replica analysis is that the average of Z^n over the Gaussian noise gives an effective Hamiltonian in which the different copies (that for a given sequence did not interact) now are explicitly interacting together

$$\langle Z^n \rangle = \int \prod_\alpha D\vec{r}_\alpha(s) \exp\left\{ -\frac{1}{2a^2} \sum_{\alpha=1}^n \int_0^N ds \left[\frac{d\vec{r}_\alpha(s)}{ds} \right]^2 + \frac{\beta^2 \langle v^2 \rangle}{2} \times \sum_{\alpha,\beta} \int_0^N ds \int_0^N ds' \delta[r_\alpha(s) - r_\alpha(s')] \delta[r_\beta(s) - r_\beta(s')] \right\}, \quad 16.$$

where we have ignored the effects responsible for chain collapse, and $\langle v^2 \rangle$ is the mean square value of the random variables v_{ij} values. The effective interaction between copies simply reflects the fact that a state that is low energy for one copy of the system will also be a low-energy state for the other; thus they will seem to be attracted to the same point in configuration space and effectively seem to be attracted to each other. An interesting analogy is to children in the

presence of an ice cream wagon. Each child is attracted to the ice cream wagon because of its desirable product. Because of this, when many children hear the ice cream wagon, they all go to the same place as if they were attracted to each other, but instead, they are simply all finding a common favorable situation. Eventually the replica interaction can evolve into a real one. (Note how lasting friendships were made by those kids.) The apparent attraction between copies allows one to introduce a new order parameter, reflecting the similarity of different thermally occupied configurations. This order parameter is analogous to one already introduced in the theory of spin glasses by Edwards & Anderson (92). It has the form

$$q_{\alpha\beta}(r, r') = \int_0^N ds \delta\left[\vec{r}_\alpha(s) - \vec{r}\right] \delta\left[\vec{r}_\beta(s) - \vec{r}'\right]. \qquad 17.$$

We can see that this order parameter measures whether two different copies of the same system have related three-dimensional structures. In their first treatment of the problem, Garel & Orland (93) assumed that $q_{\alpha\beta}$ was symmetrical. While this crudely locates the phase transition in temperature, the symmetrical approximation is not good enough for describing the nature of the transition. As Garel & Orland themselves argued, the RHP should have a phase transition something like that of the Potts spin glass (93). A Potts spin glass is a random magnetic system in which the spins have multiple pointing directions. They tried to make this connection through thinking about the lattice version of the theory rather than the continuum path integral. In fact, there are more general grounds for believing this. While seemingly exotic, Potts spin glasses represent the general case of a system of interacting objects lacking special symmetry. Gross and coworkers showed that such systems have a phase transition in which the symmetry between the different copies of replicas is broken, but in a particularly simple way (94).

Later, Kirkpatrick & Wolynes showed that this type of replica symmetry breaking exhibited by Potts glasses corresponds to an entropy crisis essentially like that in the random energy model (95), the only difference being the configurational entropy of groups of states or basins of attractions located at the transition point. Thus microscopically, the coarse-grained energy landscape would resemble that of the REM used earlier. Garel & Orland (93) did not explicitly construct solutions for the path integral that lacked replica symmetry. This was first done by Shakhnovich & Gutin (91). Their replica calculations showed that it was important for the polymer globule to first collapse before the glass transition exhibiting replica symmetry breaking could occur. They were able to describe the replica symmetry breaking by mapping the problem of the polymer chains interacting with each other onto a quantum mechanical problem in $3n$-dimensional space. The important quantitative result from the solution

of this quantum mechanical problem is the scale of the vibrational fluctuations around a given structure. They were able to show that this microscopic scale has a discontinuous jump at some point. This discontinuous jump is typical of a Potts glass phase transition, thus showing explicitly in a microscopic model that the RHP leads to a coarse-grained energy landscape like that of the REM.

The replica techniques importantly allow one to address other questions that arise from the microscopic nature of the randomness in the sequence. Instead of taking the pair interactions to be individually Gaussian random, one can take each amino acid of the sequence to be a multivalued random variable. This, of course, gives rise to correlations between different pair interaction energies. Garel & Orland investigated such a model in which the interactions are of the contact type and separable into a sum of terms involving the physicochemical properties of the amino acids in the sequences that are taken to be random (89).

An interesting aspect of this model is that it can undergo a rather simple type of ordering, not involving replica symmetry breaking. In this ordering, the different sorts of residues can separate from each other. This so-called microphase separation is the essential feature of many models of folding, notably those due to Chan & Dill (61) that arise from quantitatively modeling Kauzman's seminal suggestion that hydrophobic residues find themselves invariably in the core of a protein, while the surface is largely hydrophilic (5).

Whether random sequences lead to microphase separation or to replica symmetry breaking depends on the rigidity of the polymer chain and the number of residue types. If only two types of residues are used and the chains are very flexible, microphase separation wins, whereas if many types of residues are used, or the chains are rigid, the Potts spin glass type of transition is more important. A detailed description of this competition in both phases of random heteropolymers has been carried out in a series of papers by Shakhnovich and collaborators (96–98). Pande et al (99) have also developed a general replica formalism to deal with this problem for the multiletter code.

The replica methods require a specification of the ensemble of interactions. They are thus most easy to apply to the fully random heteropolymer that does not include the constraints of minimal frustration. The replica tricks were first used to study a minimally frustrated model of protein folding by Sasai & Wolynes (100, 101). They explicitly studied the associative memory Hamiltonian introduced by Friedrichs & Wolynes for structure prediction (102). Most of their analysis, however, is based on a model in which there is a particular structure that dominates the energy landscape while competing structures contribute to a Gaussian random noise. Thus their analysis applies to a phenomenological model that is minimally frustrated but differs from lattice models only in having both short- and long-range interactions in space and sequence. The Gaussian noise, which still acts in addition to the terms in the energy function

correctly guiding the protein to its minimally frustrated ground state, is averaged over using the replica trick. Because of the long-range interactions in sequence, the direct mapping onto a simple time-independent quantum mechanical problem used by Shakhnovich & Gutin (91) was not available to Sasai & Wolynes (100, 101). They introduced a new approach to the replica statistical mechanics based on a replica version of Feynman's treatment of the polaron. The same technique was later used for other problems involving random higher-dimensional manifolds (103). Sasai & Wolynes (100, 101) introduced a reference Hamiltonian that reflects the possible phase transitions of collapse, ordering in the correct structure and the replica symmetry breaking. The familiar Peierls variation principle, now in replica space, was used to find the phase diagram. Their diagram is essentially isomorphic to the phase diagrams obtained for minimally frustrated partially random heteropolymers using random energy model approximations by Bryngelson & Wolynes (10).

An interesting feature of the replica approaches is the importance of the vibrational entropy describing fluctuations around both local minima and the native structure. Both the folding transition and the trapping transitions then have instabilities and an associated Lindemann criterion for the mean square fluctuations about the appropriate structures (100, 101). A similar such instability point occurs in the theory of the liquid glass transition as well as in the more general Potts spin glass. The instability point for the glassy metastable states reflects the point at which dynamics changes from an inactivated hydrodynamic type well described by mode-coupling theory to one involving hopping between different states. This dynamical transition occurs at a higher temperature than the ideal glass transition temperature. The chain dynamics should depend greatly on the proximity to the dynamical transition. For the associative memory protein models with long-range interactions in space studied by Sasai & Wolynes (100), the dynamical freezing temperature is close to that at which the globule is first formed. The transition to activated dynamics, however, depends on the range of the interactions. Using the same approach, Takada & Wolynes have recently discovered that models that are minimally frustrated but with short-range interactions have a distinctly lower dynamical transition temperature (104). Thus activated escape from traps actually occurs at a lower temperature. A phase diagram indicating both instability points for the folded structure and for traps, reflecting the entropic smoothing of the landscape, is shown in Figure 8.

In these microscopic treatments, the minimal frustration is still put in phenomenologically. One might well ask how to pre-specify the ensemble of foldable sequences. This may be very important when one considers recent efforts using combinatorial synthesis to make foldable proteins (105, 106). As we have already seen from phenomenological arguments, foldable sequences must be

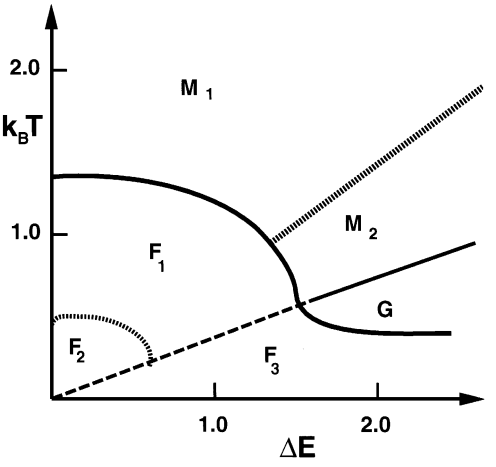

Figure 8 Phase diagram for a short-range minimally frustrated heteropolymer. Both thermodynamic transition lines and instability curves are shown. The ordinate is the temperature in units of stability gap, while the abscissa is the ruggedness in the same units. Solid curves separate different static phases, and dotted curves represent the boundaries at which metastable states become unstable and disappear. The dashed curve shows the glass transition in the metastable unfolded state. The M_1 region is the equilibrium molten globule phase where traps are entropically unstable, yielding a monotonous free energy landscape. The M_2 region still corresponds to the molten globule state but where the landscape is rugged because the system is below the dynamical glass transition T_A. G corresponds to the ideal thermodynamic glassy phase. In the F_1 region the protein is folded but a free-energy barrier exists to the unfolded state while in F_2 folding is downhill. F_3 corresponds to a region where the folded state is stable but, since the molten globule is glassy folding, is very slow. The values of the parameters used in the calculations by Takada & Wolynes (104) were made to correspond roughly to the lattice simulations described in the next section.

stable at a higher temperature than the glass transition for most random sequences. This can be turned into a formal criterion for describing the ensemble of foldable sequences by specifying the energy of the ground state structure. Ramanathan & Shakhnovich have introduced an ensemble of selected sequences according to a selection temperature that is thermodynamically conjugate to the ground state energy (107). This ensemble, rather than the purely random one, can now be used along with the explicit replica technique. The phase diagram obtained in this evolutionarily selected ensemble agrees with that of Sasai & Wolynes (100) in qualitative form, although the quantitative mapping between the two models is not entirely trivial.

An alternate and very interesting approach to an ensemble of minimally frustrated proteins has been proposed by Pande et al (108). They invented an ensemble describing a polymer that is meant to bind to a particular ligand. This

also gives rise to minimally frustrated sequences. Their mathematical ensemble corresponds rather closely to the experimental procedure of imprinting. In this method, sequences are assembled on a template and chemically bonded together. This might be described as a Lamarckian evolution of foldability as opposed to the Darwinian approach of the Ramanathan-Shakhnovich ensemble. Various scenarios for the origin of foldable proteins in the primeval soup can be based on such imprinting mechanisms. It resembles the famous clay origin of life espoused by Cairns-Smith (109).

The statistical mechanics of sequence selection can explain many provocative features of the natural taxonomy of proteins. Finkelstein et al use it to understand the paucity of energetically costly local structures in natural proteins, although it would seem these could be allowed by compensating stabilization over design of other parts of the protein (110, 111). Wolynes has used it to discuss the relation between the observed approximate symmetries of natural proteins and the symmetry of clusters with magic numbers of atoms (112), an issue also addressed by lattice simulations and other arguments (113, 114).

Almost all microscopic approaches to the energy landscape of proteins and heteropolymers rely on approximations with an element of mean field theory. More needs to be done to address this. Shakhnovich (97) has deduced a Ginzburg criterion suggesting that mean field theory is exact for equilibrium questions. This analysis is based on the usual low-order perturbation corrections to mean field (97). While adequate for the usual polymer aspects, it does not include the nonperturbative effects that are known to be important in spin glasses (95, 115–118). Thirumalai has presented a qualitative argument that contains such effects (119). Franz et al have argued that the mean field approach is inadequate because one can apply the same approximation to the random directed-polymer, which models a noninteracting polymer absorbed on a random surface, but this has been shown to be mean field–like only in very high dimensions (120). The spin glass defect configurations, which are worrying for RHP, are even more effective in the directed-polymer case because of its special quasi one-dimensionality of interactions along the sequence. At this point the mean field arguments seem adequate for the smaller protein-like systems studied by simulations. This may be because the defects are large, i.e. of the order of the size of natural protein domains.

SIMULATION OF MINIMALIST MODELS

As discussed above, the energy landscape theory suggests there are several scenarios for folding kinetics and mechanisms (15). This diversity has been observed in many computer simulations of the folding event and has aroused considerable interest in the folding community (121). Such simulations can be

carried at various levels. Ideally we would like to have all calculations performed at the atomistic level, including all the details of the protein-solvent environment. The main limitation is that the overall time scale of the folding events can extend from milliseconds to seconds, far from the reach of the present molecular dynamics simulations. For this reason, this approach has primarily provided insights into local aspects of folding (122–125) (such as helix formation) and has also been successful in characterizing ensembles of deliberately unfolded proteins (126–132). Because of the limitations of the atomistic models, more has been learned about the overall folding process from minimalist models of protein folding (133). They include the simple lattice models exploited early on by Gō (134) and Covell, Jernigan, and coworkers (135–138) and developed most extensively by Dill and coworkers (16, 61, 139, 140). More recently, several other groups, including Shakhnovich, Thirumalai, Karplus, Scheraga, Pande and Grosberg, Socci, and Onuchic (22, 54, 141–148) have vigorously pursued the simple forms of these models toward a better understanding of the folding mechanism. Lattice models have been elaborated quite completely by Skolnick and coworkers (149–153) into schemes that show great promise for structure prediction. Minimalist models are not confined to lattices. Off-lattice models were used very early on by Levitt & Warshel (154) and more recently by Friedrichs, Wolynes, and coworkers (102, 155, 156), Thirumalai and collaborators (157–160), Berry, Wales, and collaborators (161–164), Sasai (165), and Friesner, Honig, and collaborators (166, 167). Many of the same features discussed by simulators of lattice models have been seen as well in the continuum, sometimes earlier in fact. Because of the greater variety of continuum models, there has been less quantitative comparison of simulation results to analytical theory, which is why we focus here on the basics of lattice models and their interpretation using landscape ideas.

The advantage of studying the lattice models is that an in-depth analysis (on occasion both exhaustive and exhausting) can be performed, yielding detailed answers and information. Interpreting the results using the landscape theory is essential because minimalist models make many artificial simplifications, and theory should be the guide to what are artifacts of the model and what are common features of model proteins and real ones. It is important to bear in mind the simplifications associated with these lattice models. To make this point clear we now discuss the basic protocol of lattice simulations, using as an example some of the studies of Socci & Onuchic.

The most studied lattice model consists of a string of connected beads with fixed bond length that are allowed to move on a cubic lattice in two or three dimensions. This has a venerable history in polymer physics, and its use for proteins was pioneered by Gō and coworkers (134). The length of the chain is at the simulator's disposal, but in three-dimensions, the paradigm system is a

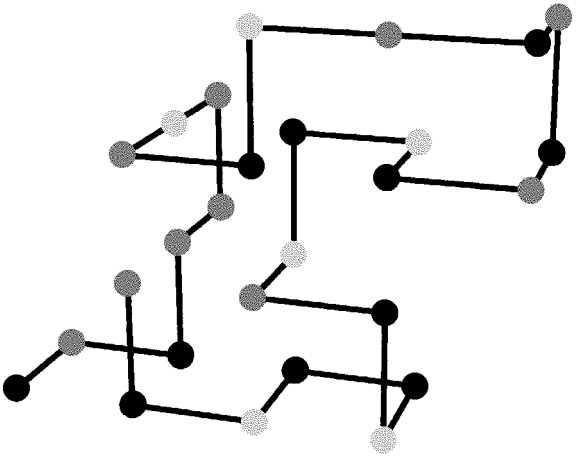

Figure 9 A random configuration for a 27-mer sequence encountered en route to the native structure. The maximally compacted conformation for this sequence, not shown here, has the shape of a $3 \times 3 \times 3$ cube.

27-mer. A felicitous aspect of the 27-mer is that, in addition to stochastic sampling methods, one is able to completely enumerate all its maximally compact structures ($3 \times 3 \times 3$ cubes) (140). In order to address the sequence-to-structure information transfer problem, which is the central question of folding, at least two different kinds of beads are needed. Figure 9 shows a random configuration encountered en route to the native structure. Kinetic simulations start from a random configuration on the lattice, and a series of configurations are generated by the conventional Monte Carlo kinetics procedure (168–171). The relationship between lattice dynamics and real time dynamics has been a main concern since the early 1970s. A variety of move sets has been investigated both in the context of proteins and for dense phases of polymers. In the dense polymer phase, there is a strong dependence of the dynamics on the move set chosen, but for model proteins, which have a large surface area allowing less constrained motion, the effect of changing the move set is usually simply a rescaling of the overall time scale. Global moves are not normally allowed, so for larger proteins where domain motion may be involved, the move set dependence may become more critical.

Using the simplest possible potentials between the different beads already yields interesting information. The first set of potentials (61, 134) only included attractive interactions between hydrophobic groups (H) that are neighbors in the lattice. A slightly different choice of simple potentials has also been widely used.

Residues attractively interact with nearest neighbors in the lattice with a stronger attraction when the residues are of the same kind. Potentials mimicking the full variety of amino acids are also commonly employed (136, 151, 172, 173), and occasionally pair interactions are chosen from independent random distributions, which can facilitate statistical analysis and comparison with analytical theory (144).

To get at the kinetics, a series of runs are initiated. Statistics for the times to achieve different ensembles of configurations are monitored. The mean first passage times are particularly enlightening. Figure 10 shows folding and collapse times as a function of temperature for a series of sequences with two and three kinds of beads (15, 23, 148, 174). The nonmonotonic temperature dependence is striking and very much in keeping with the behavior expected from analytical energy landscape theory, where folding down the funnel is accelerated by a stronger thermodynamical driving force as the temperature decreases, but transient trapping in local minima on the rugged energy landscape is enhanced with decreasing temperature. For this potential the time to form collapsed structures is not strongly dependent on sequence at high to moderate temperatures but becomes strongly sequence dependent at a temperature near the theoretically expected glass transition temperature. On the other hand, folding is strongly sequence dependent, even at high temperatures. The fastest folders are the ones with the largest stability gap. This correlation is most dramatic when we consider folding under conditions where the folded state is required to be stable. When $T_F < T_G$, the rate of folding is more than 10,000 times slower (limit of the simulation time) than the good folding sequences. The fast folders that correspond to well-designed sequences exhibit exponential folding kinetics around T_F and become nonexponential as T_G is approached. This distinction between good and bad folders becomes more important as the systems grow in size. The ratio between T_F and T_G becomes more kinetically relevant for larger systems. As system size increases, since the configurational space grows exponentially, a larger separation between T_F and T_G becomes necessary for a single domain protein. Larger proteins in nature are generally multidomain.

Many minimalist simulations have addressed the question of foldability. As discussed earlier, this is a simplified formulation, since we must specify necessary time scales and thermodynamic conditions to be precise. In any case, most of them come to conclusions that are consistent with the T_F/T_G criterion that quantifies the minimal frustration principle to achieve fast stable folding proteins. Two early off-lattice simulations make this connection. Friederichs, Wolynes, and coworkers discuss the capacity of associative-memory Hamiltonians to fold heteropolymer chains to correct structures in terms of this ratio (102, 155, 156). Thirumalai and coworkers designed an off-lattice bead model and examined its lower-energy states (157). They suggested the foldability

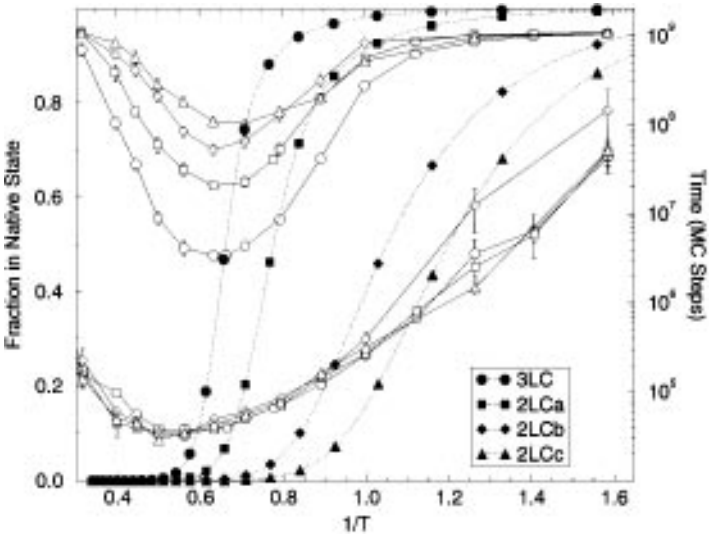

Figure 10 Folding times, collapse times, and thermodynamic stability versus the inverse temperature for several designed two-letter code (2LC) and three-letter code (3LC) 27-mers. The potential used favors strong (weak) attractive interaction between neighboring residues of the same (different) kind. Since the entire range of temperatures is scanned, there is only one free parameter that determines how fast folding is compared to collapse for the well-designed sequences. The results presented here are for the potentials where collapse is rapid relative to folding. The units of energy are arbitrary and are chosen to have a theoretical glass temperature close to one. All the two-letter code sequences have been designed under the constraint of a ratio of 14:13 between monomer types. The three-letter code sequences are designed by selectively introducing a third kind of residue in our designed two-letter code sequences (23, 54, 148, 174).

The (kinetic) glass transition is chosen when folding times become extremely long (of the order of 100 times longer than the fastest folding time) (see 23 for details). The mean first passage time for collapse is sequence independent, until glassy dynamics takes over for $T \sim 1$. (The maximally compact conformations have 28 contacts, and we define collapse when the first 25 contacts are made.)

The ground state energies for sequences 2LCa, 2LCb, and 2LCc are -84, -80, and -76, respectively. For two-letter sequences with fixed composition, the one with the lowest ground state seems to be the fastest folder. This difference becomes pronounced upon examination of the folding temperature, T_F where the probability of being in the native configuration is 50%. The fast sequences have $T_F > T_G$. From the 2LC sequences, it appears that the solution to the design problem would be merely to minimize the energy of the native configuration. This misconception is clarified upon comparison of the 2LC and 3LC sequences. The best designed 2LC sequence has $T_F/T_G \sim 1.3$. By changing the designed 2LC sequences with a third kind of monomer, the ground state energy of the best designed sequences did not change, but the energies of the other collapsed configurations were raised. This effectively increased the stability gap and thus T_F, and since these nonnative collapsed states have weaker nonnative contacts, T_G became smaller. With the help of the third monomer it was therefore possible to raise T_F/T_G to 1.6. The design strategy should be to maximize the energy gap between the native configuration and a typical collapsed configuration (stability gap) in units of the roughness.

was related to a gap in the low-energy spectrum that would be expected if the design satisfied a large T_F/T_G. Leopold et al have shown that good folding sequences have a funnel-like landscape that only exists for minimally frustrated heteropolymers and illustrated both the spectrum of states and kinetic connectivities of a folding and nonfolding sequence (22). While preliminary results of another lattice calculation suggested no connection of foldability and minimal frustration (141), a later extension of these studies by Sali et al (144) concurs with the landscape ideas. They also formulated their foldability criterion in terms of an energy gap. In a survey of two hundred random sequences of 27-mer, kinetic foldability at a fixed temperature correlates well with the gap for this family of sequences. They did not account for stability when analyzing their data. It is clear when this is taken into account that the appropriate gap involves the difference between the folded state energy and the thermal average of typical collapsed unfolded states with which it competes. That gap is more precisely related to T_F/T_G for a family of sequences with similar roughness, as can be noticed for the collection of 2LC sequences in Figure 10 (54, 148). Simulations on continuum models by Hao & Scheraga also provide supporting evidence for the T_F/T_G criterion (175, 176).

Joint consideration of stability and kinetics is vital for understanding foldability. The fact that several studies only focused on the kinetic aspects has been the source of most of the controversies. Figure 10 clarifies this point. It shows that sequences that have mean first passage folding times not very different for a given high temperature differ substantially in their ability to fold if compared at their respective T_F.

Recently, renewed controversy has arisen about quantifying foldability. Klimov & Thirumalai suggested that one should compare T_F with the collapse temperature (177). Their simulation shows that the larger this ratio, the faster sequences fold. This criterion is not inconsistent with the T_F/T_G criterion, since Bryngelson & Wolynes showed the T_G is always below the collapse temperature (62). For fixed average hydrophobicity, T_G and the collapse temperature are strongly correlated (54, 62). When T_G of the collapsed manifold is fixed, decreasing the average hydrophobicity does in general speed up folding as expected, since less compact structures yield less ruggedness (54). An amusing feature of the Klimov & Thirumalai (177) work is that they clearly illustrated that the gap must be defined through the stability gap defined earlier (23), not by the first excited state gap, with which it is sometimes confused. Their study shows no correlation between stability and the first excited state gap.

Other routes to minimally frustrated sequences have been explored in lattice studies. These include sequence design on the lattice by Shakhnovich & Gutin (142) and Pande et al (147) using explicitly the minimal frustration strategies. Banavar and coworkers (178) investigated sequence design using a hierarchical

picture of sequence organization that was not overtly based on the minimal frustration principle. After examining the energy spectra of their sequences, they discovered again consistency with the T_F/T_G criterion (178).

Simulations can also be compared more quantitatively to energy landscape theory (174). A protein folding along the funnel shown in Figure 4 moves through an ensemble of partially ordered structures characterized by the similarity measure Q (Q is the fraction of native contacts). In Figure 11, we show a trajectory of the Q and A coordinates superimposed on the free-energy contours for these collective variables (A is the fraction of native angles). The motion is very erratic and looks Brownian. A simple description of the overall dynamics within the folding funnel thus should be obtained using a diffusion equation.

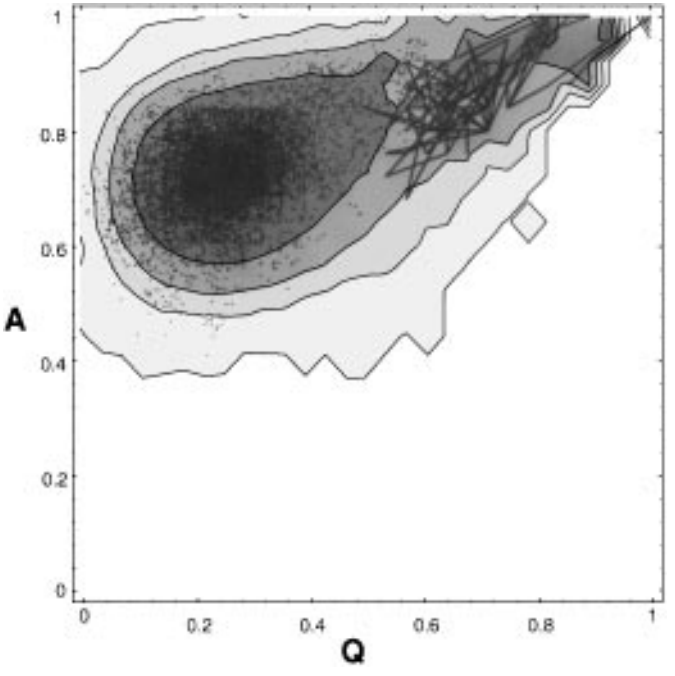

Figure 11 A transition trajectory projected onto the *Q*-*A* plane for the designed three-letter code sequence (see Figure 10). The time span is approximately 25% of the folding time, which is $\sim 3 \times 10^6$ Monte Carlo steps. The last part of the trajectory that is connected by lines includes only one eighth of the full trajectory. The free-energy contours are from -67.5 to -82.5 in increments of 2.5 in energy units such that $k_B T_F \approx 1.5$. The trajectory clearly indicates the diffusive nature of the Q and A dynamics, and that the transition event over the barrier is very short compared to the full folding time. See Reference 23 for more details.

The gradient of the free energy determines the instantaneous drift velocity down the funnel. Superimposed on the drift are stochastic fluctuations in Q reflecting individual escapes from traps. Even though this diffusion picture appears to be a continuous process, the existence of a barrier in the free-energy profile leads to exponential folding kinetics. The roughness of the energy landscape at any stage acts like a set of speed bumps slowing both the drift and the superimposed Brownian movement. At a given temperature the population of the various structural strata changes with time according to

$$\frac{\partial P(Q,t)}{\partial t} = \frac{\partial}{\partial Q}\left\{ D(Q,T)\left[\frac{\partial P(Q,t)}{\partial Q} + P(Q,t)\frac{\partial \beta F(Q,T)}{\partial Q}\right]\right\}. \qquad 18.$$

In general the local configurational diffusion coefficient D depends on the roughness of the energy surface, which determines the escape time from traps. The diffusion coefficient is inversely proportional to the lifetime $\tau(Q)$ of a microstate with similarity Q to the native state. If the microstate is deep, it will be long-lived and the diffusion coefficient becomes small. In the REM analysis, being trapped in this microstate characterized by a roughness $\Delta E(Q)^2$, motion must take place over an energy barrier $\bar{E}(Q) - E_{\text{m.p.}}(Q) = \Delta E^2(Q)/k_{\text{B}}T$ in the time τ_o it takes for a large segment of the chain to move. This gives an escape time from the local traps that is super-Arrhenius

$$\tau(Q) = \tau_o \exp\left[\Delta E^2(Q)/(k_{\text{B}}T)^2\right]. \qquad 19.$$

While the specific temperature dependence is now subject to much discussion, this relation can be at least used as a phenomenological one. For a well-designed 3LC sequence (see Figure 10), this diffusion coefficient can be obtained directly by measuring the correlation function of the fluctuations of Q from simulation data of the collapsed states alone.

In the case of fast downhill folding at a fixed temperature shown in a Type 0 scenario, a kinetic folding bottleneck occurs at a region $Q^{\ddagger}_{\text{kin}}$ with the maximum lifetime or the smallest diffusion coefficient. This maximum lifetime is also a simple estimate of the overall folding time

$$\tau_{\text{f}} \approx \tau_{\max}(Q^{\ddagger}_{\text{kin}}). \qquad 20.$$

For a bistable system as in Type I and Type IIA scenarios, the overall folding time τ_{f} will be determined by the difficulty to overcome the free-energy barrier and a prefactor that depends on the ruggedness of the energy landscape

$$\tau_{\text{f}} \approx \left\langle \Delta Q^2_{\text{MG}} \right\rangle D^{-1}(Q^{\ddagger}) e^{\Delta F^{\ddagger}/k_{\text{B}}T}, \qquad 21.$$

where $\Delta F_{\ddagger}$ is the free-energy barrier measured from the unfolded minimum to the top of the thermodynamic barrier. $\langle \Delta Q^2_{\text{MG}} \rangle$ is the mean square fluctuation

of the configuration coordinate in the molten globule state. This equation suggests that an Arrhenius plot of folding time versus inverse temperature would be curved, and such behavior is frequently observed in protein folding experiments (179) and lattice simulations, as seen in Figure 10. As the temperature is decreased, the escape time will increase until the local glass transition temperature $T_G(Q)$ is reached. For $T < T_G(Q)$, the protein has kinetic access to very few structures, and the protein is effectively frozen into a single or several low-energy states, each with a specific escape rate. A traditional experimentalist would be tempted to describe escape from these traps as a pathway. In this case the kinetics are dominated by the details of the specific landscape, and the expressions for the folding time and the diffusion coefficient would have to be modified for a quantitative treatment.

The theoretical ideas above can be used for quantitatively predicting folding times in model proteins with a realistic energy landscape topography. Socci et al (174) have shown that as long as the glass transition falls after the transition region (top of the barrier in the free energy profile for the collective reaction coordinate) this is the case. In this limit the single dominant funnel picture is appropriate. The system they studied is the best designed 3LC 27-mer lattice model (Figure 10). Figure 11 shows a folding trajectory of the Q coordinate superimposed on a plot of the free energy. Most of the trajectory consists of diffusive motion about the molten globule region. Once the barrier has been surmounted, folding occurs rapidly, taking roughly 10^5 Monte Carlo steps ($\approx .03\tau_f$).

From Equation 21 it becomes clear that to estimate the folding times at a variety of temperatures, knowledge of the free-energy barrier alone is insufficient. Information about the dynamics must be obtained by calculating the configurational diffusion coefficient through the complex energy landscape. In general the diffusion coefficient will depend on Q, but one more simplification is assumed here. Only the average value of D, computed for states in the molten globule band, is inferred from simulations. This was done by computing the correlation function of the fluctuations of the reaction coordinate $\Delta Q(t)$.

With the diffusion coefficients and free-energy surfaces in hand, the analytical predictions given by Equation 21 were tested. The results are presented in Figure 12, and the agreement between theory and simulations is remarkable. Thus we see the analytical theory based on the actual molten globule dynamics and funnel free-energy profile is not simply qualitatively correct but can be used for quantitative predictions of the folding time over a wide thermodynamic range, at least for well-designed sequences at temperatures above the glass transition.

To explore this correspondence between real proteins and minimalist protein folding models, the effect of the additional degrees of freedom have to be included, particularly secondary structure possessed by real proteins. If separate

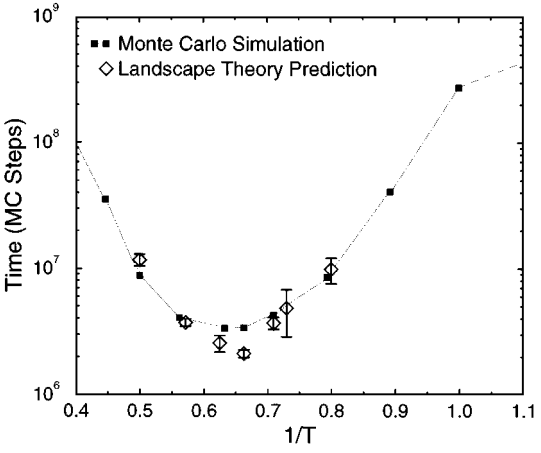

Figure 12 Comparison of the mean first passage times from Monte Carlo simulations with the theoretical predictions using the landscape theory by Socci et al (174) for the three-letter code 27-mer sequence presented in Figure 10. The agreement between theory and experiments is outstanding (rates differ by less than a factor of two), clearly indicating that the folding kinetics can be represented as a diffusive event in the Q reaction coordinate. The configurational diffusion coefficient is obtained by computing the correlation function of the fluctuations of the reaction coordinate $\Delta Q(t)$ only for states in the molten globule band (collapsed states only).

phase transitions for ordering these additional degrees of freedom intervene during folding, a multistep mechanism could result, but in a major part of the complex phase diagram, the effect of these extra degrees of freedom is to renormalize the entropy and energy scales for the protein folding funnel (see phase diagram in the previous section). In the regime of fast collapse, a simple version of this renormalization can be performed analytically, and a law of corresponding states relates the 27-mer lattice model to a 60–amino acid helical protein. This law of corresponding states is in a later section of this review.

ADVANCED DYNAMICS ISSUES

The analysis of folding kinetics using a single diffusive reaction or progress coordinate is at present the approach that has been most worked out quantitatively. Another description, which is not orthogonal, has also received attention. When a barrier exists, one may describe folding down a funnel as being nucleation in a finite size system (10, 13, 14, 23, 62, 180). If the nucleus is specific and very small, the quantitative use of a global reaction coordinate for kinetics can be misleading. To understand the magnitude of the effects, the scientific question

becomes How big and how localized is the nucleus? If it is large or delocalized, single progress coordinate language will be more than adequate, with structural inhomogeneities being a significant but secondary perturbation. This issue depends on details of the microscopic potentials and the thermodynamic conditions of folding. The uniqueness of the nucleus depends on the heterogeneity of contact energy and thus the quality of design of the protein. Classical nucleation theory suggests nucleus size further depends on surface tension and the driving force to the ground state structure, both of which depend on the thermodynamic conditions (181, 182). In 1990, using the classical theory, Bryngelson & Wolynes addressed the nucleus size issue, and suggested that the nucleus was quite large, comparable in size to the whole protein (62). Elegantly refining this argument to take into account the shape of the protein, Finkelstein suggested recently that the nucleus would be somewhat smaller, about one third the size of a single domain protein (183). However, simulations by Shakhnovich's group led them to suggest that the nucleus for their designed lattice protein is much smaller, perhaps containing only three or four key residues (143). They further showed that these key residues are at locations in the lattice protein where it is most constrained and argue that naturally conserved residues in real protein are likely to be this specific nucleus (172). Indeed, experimentally, Fersht and coworkers have beautifully shown in the most studied example, chymotrypsin inhibitor CI2, that different residues participate in the folding transition state ensemble to varying extents (35). The most significant residues are indeed highly conserved, but a closer examination of the experimental data suggests that the participation of residues in the transition is still highly delocalized, although it does not encompass more than half of the protein.

The specificity of the nucleus depends on the heterogeneity of the contact energies and on the entropy losses of forming contacts during folding. When heterogeneity is small the delocalization is significant. Socci and coworkers showed that for a well designed 27-mer whose properties resemble fast folding proteins, there is a wide but unimodal participation distribution in the folding transition state that qualitatively resembles the histogram of CI2 experimental data (180). A bimodal distribution can be found for larger or more poorly designed models.

Shoemaker et al characterized the structural correlations in the transition state ensemble of CI2 by using a mean field theory that takes into account the specific energies of making pair contacts; but they still used a single progress coordinate for folding (184). The calculation agrees well in locating the central core of folding participation and exhibits a good correlation with experimental results from protein engineering for fractional participation for buried residues but not a very good correlation with highly solvent exposed residues that were inadequate for the energy function used.

Energetic heterogeneity also can lead to a breakup of folding into kinetic domains (62, 185). In lattice studies this has been observed by Shakhnovich and coworkers (18, 143, 186). One sequence with a specific nucleus in an early Shakhnovich study (143) actually exhibits two such kinetic domains. Panchenko et al have used empirical energy functions along with energy landscape analysis to scan for such kinetic subdomains, which they call foldons in a large group of natural proteins (187). The foldons they find are comparable in size to the exons found in the DNA sequences for these proteins but seem not to be always identical to them. This may be relevant to the debate about why genes have pieces. There is some overlap of the predicted foldons with late-stage kinetic intermediates in cases where this has been studied. Most notable here is the comparison with the late-stage folding of lysozyme and lactoalbumin studied by Dobson and coworkers (188, 189).

Another constraint on the the use of the reaction coordinate idea is that a sense of locality of moves must be imposed on the kinetic connectivity. This may be hard to satisfy, especially if there are topological constraints leading to dead ends. An alternative formulation of rates can be modeled with global but random connectivities. Such globally connected models for minimally frustrated landscapes, both with and without correlation, have been studied and produce behavior for the variation of rates with thermodynamics that is qualitatively similar to that of diffusive dynamics (190, 191). Zwanzig has shown that the kinetic description depends weakly on the connectivity if it is sufficiently large (192). Wang et al have gone further and discussed the relationship between the uniqueness of kinetic paths and connectivity (193).

Equally as important as the thermodynamic free-energy profile along the reaction coordinate to understand the folding rate is the configurational diffusion coefficient, which reflects the ease of sampling new configurations and escaping from transient trap structures. Within the REM approximation this reflects the full energetic ruggedness of the landscape. The simulations of Socci et al show that the configurational diffusion coefficient of the 27-mer can be fit with the REM form but with a diminished apparent ruggedness or with an Arrhenius law with a large barrier (174). This suggests a substantial but partial participation of the chain in the motion leading to trap escape. A similar diminished barrier was also found at low temperature by Gutin and collaborators (194). Those studies also show trap escape barriers increase with system size at least at low temperatures.

Analytical treatments of kinetics that go beyond the REM and include the correlated nature of the landscape are beginning to be developed. Takada & Wolynes used replica techniques developed for mean field spin glasses to estimate the barriers between the onset of activated dynamics (T_A) and the glass transition at T_K (104). The predicted barrier heights at T_K are in harmony with the

low-temperature simulation. The generalized REM also can be used to describe global escape from traps. Wang et al (193) showed that the GREM exhibits the two transitions at T_A and T_K and that the correlations significantly diminish the barrier heights. The dynamic GREM approximation shows a smaller temperature dependence than that of the simulation values (195).

The existence of a dynamic transition (T_A) is a bit controversial. We must recognize that it can only be crisp for large systems with long-range interactions. Takada & Wolynes have generalized their earlier mean field calculation to allow entropic droplet configurations as escape routes (196). The most important droplets in this calculation are again similar to the entire small protein in size, suggesting the validity of a mean field approach as a starting point.

At temperatures above T_A it is not appropriate to describe chain molecule motion as activated trap escape. Instead, chain dynamics is Rouse-Zimm–like modified by mode-coupling effects. A fairly advanced mode-coupling theory for homopolymers was put forward by Schweizer (197). For the heteropolymer a variety of different mode-coupling calculations have been carried out with results at variance with each other. Roan & Shakhnovich concluded that there is no dynamic transition (198), while Thirumalai and collaborators, for a somewhat different heteropolymer model, obtained a transition that depends on chain length (199). Dawson and coworkers have carried out a numerical treatment of collapsed dynamics which shows a glass transition (200). Takada et al recently developed a mode-coupling calculation that is compatible with the replica theory results (201).

CONNECTION TO FAST FOLDING PROTEINS EXPERIMENTS

Onuchic et al developed a law of corresponding states to relate simulations of small lattice models to real proteins (23). The correspondence analysis made use of an analytical theory of helix formation in collapsed polymers that related the configurational entropy S_o to the amount of helical structure (67). The hypothesis behind this approach is that as the protein goes through the hydrophobic collapse, around 60% of helix formation has occurred (not necessarily the native helices), just as some think occurs at equilibrium globules, and this collapse is fast relative to folding. The phase diagram envisioned here is essentially shown in Figure 13. Even though both of these hypotheses, need of fast collapse and a fixed amount of secondary structure for collapsed states, are probably too restrictive, the general features of the law of corresponding states should remain the same for more elaborate approaches, at least in the lower part of the funnel. For a 60–amino acid chain at 60% helicity, $S_o \approx 40k_B$, which corresponds approximately to a conformational entropy of $0.6k_B$ per monomer unit.

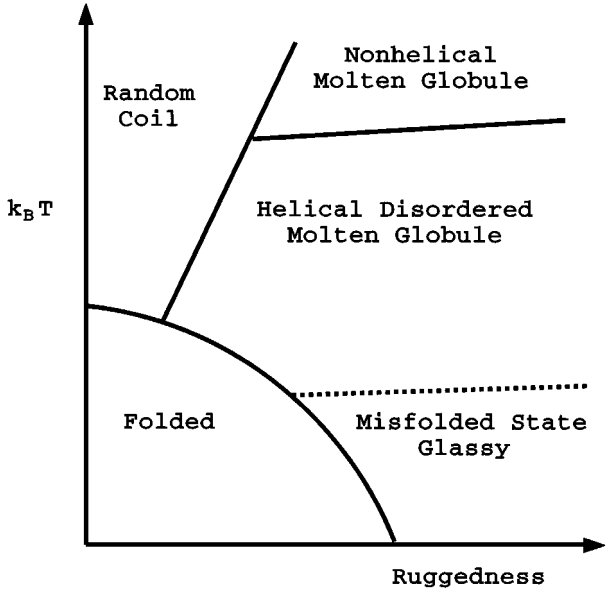

Figure 13 A possible phase diagram for a minimally frustrated, hydrophobic, helical protein. Depending on the hydrogen-bond strength, the molten globule can have either no well developed secondary structure, or may be a helical liquid crystaline molten globule, as discussed by Luthey-Schulten et al (67). The quantitative funnel shown in Figure 4 was drawn assuming the latter is encountered before the rate-limiting step of folding.

After this renormalization, a correspondence between the lattice model and real, small proteins can be done by three parameters: configurational entropy, energy ruggedness of the landscape, and the stability gap that quantifies the specificity of the native contacts. For a given value of the configurational entropy, the competition between the stability gap and the ruggedness can be quantified by the ratio T_F/T_G. T_F is the folding temperature (when the occupation of the folded state is 50%). T_G is the glass temperature, below which trapping dominates the folding event. Onuchic et al (23) estimated that a realistic folding funnel for fast folding a 60–amino acid helical protein should have a $T_F/T_G \sim 1.6$, corresponding to the designed three-letter code sequence in the lattice. The schematic folding funnel for such a protein is shown in Figure 4. At T_F folding proceeds via a Type IIB scenario with the transition state at $Q = 0.60$ and a glass transition at $Q = 0.71$ (recall that Q measures the density of native tertiary contacts).

Although the detailed funnel topography was originally proposed using the correspondence with the lattice simulations, this theoretical prediction parallels

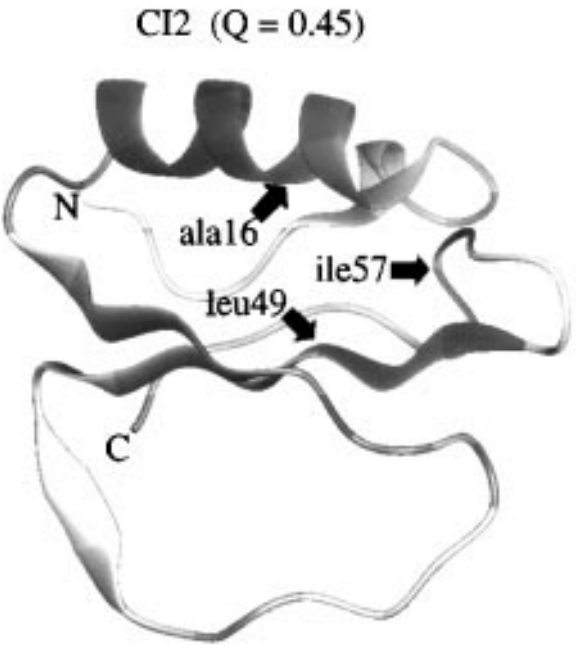

Figure 14 The folded structure of chymotrypsin inhibtor 2 (CI2) is shown for ensembles with a transition state value of $Q^* = 0.45$ and with residues colored according to their involvment in the ensemble (184). Darkest shading indicates strongest participation. The specific hot sites identified by experiment (35, 202) are indicated by arrows.

later experimental work on small helical proteins obtained by Huang & Oas (24) for the monomeric λ-repressor, a ~ 70 residue protein with largely helical structure. A transition state located near $Q = 0.5$ was obtained. Similar behavior has been observed by Fersht and collaborators (202) for chymotrypsin inhibtor 2 (CI2), a small 64-residue protein (Figure 14). Even though the comparison is not as direct because this protein is composed of both α-helices and β-sheets, again, the bottleneck is midway between the folded and unfolded regions. Slow folding proteins, whose late stages were studied earlier, usually have much higher Q transition state. Pascher et al (40) have initiated the folding of cytochrome *c* by means of photochemical electron transfer, with a strategy analogous to an earlier experiment using CO photodissociation (37). Since horse cytochrome *c* has 104 residues, its folding funnel has a different topography from the one for the smaller protein. Since the protein has more subunits, entropy is lost more rapidly in the initial stages, moving the bottleneck higher in the funnel. This collection of data clearly indicates that the transition state ensemble is

midway between folded and unfolded states and has considerable configurational entropy; it is thus composed of a large ensemble of states, much as in the corresponding picture of the funnel.

Even though native contacts participate on average with a probability of 60% in the transition state, the polymeric nature of the protein chain creates inhomogeneity for the probability of individual contacts. By sampling all configurations with $Q = 0.6$, the participation probability of individual tertiary contacts at the transition state can be computed (180). A broad distribution centered around 0.5 is found for these participation probabilities, but with values ranging from almost zero to one, indicating that the transition state bottleneck is composed by an ensemble of delocalized nuclei. A similar broad distribution for tertiary contact participation in the transition state has been observed by Boczko & Brooks in their atomistic simulations of a three-helix bundle (126). Also, experiments on CI2 by Fersht's group (35, 203) and on ChY by Serrano's group (36) provide evidence for this ensemble of delocalized nuclei in the transition state. These experiments encourage the application of landscape theory and simulations as quantitative tools.

The quantitative analysis of the folding landscape topography is just beginning. The advances in theory discussed above as well as new measurements of residual structure in molten globules and its dynamics already lead to a revision of the detailed numbers. Nevertheless, the framework is a useful one for combining results of many experiments into a coherent whole.

PROTEIN STRUCTURE PREDICTION AND DESIGN USING ENERGY LANDSCAPE IDEAS

In the previous sections we have shown how the energy landscape theory can be used to interpret the kinetics and thermodynamics of protein folding. It has long been the hope that such physicochemical work would contribute to the practical aspects of predicting the structure of a protein from its sequence and of the design of novel proteins from scratch. These two problems are often approached in an artistic way that, while pleasing and sometimes successful, lacks a framework for making organized progress. Energy landscape ideas like those used to understand protein physical chemistry provide a framework for these practical engineering questions. The essence of landscape engineering is that in its simplest form the theory requires quantifying only a few parameters: δE_s, the stability gap between the ground (native or folded) state of the protein and the mean of the excited (misfolded) states and ΔE, the roughness of the energy landscape. Nature has designed its fastest folding proteins to maximize the dimensionless ratio, $\frac{T_F}{T_G} \approx \frac{\delta E_s}{\Delta E} \cdot \sqrt{\frac{2k_B}{S_o}}$, and this automatically suggests a

physically meaningful criterion in the design of energy functions for protein folding in machina. The essential feature of a potential function is that the energy of a sequence in its native structure measured from the mean of the misfolded states is much larger than the width of that distribution. In such a case the folded conformation will automatically have an energy lower than that of all alternative conformations. Nearly all algorithms for predicting structures rely at some point on minimizing an energy function for a given sequence. While the algorithm for minimization may not be slavishly identical to the physical motions, most algorithms still have the property that their computational speed or reliability depends on the discrimination between the energy of the correct folded conformation and other false alternatives. Therefore, effective potential functions for structure prediction should satisfy the principle of minimal frustration just as they do in nature. Similarly, to design a new sequence that will kinetically and thermodynamically fold, the principle of minimal frustration must be respected if robust results are desired. Otherwise, small design flaws will not be tolerated (13, 14).

The energy functions used in structure prediction fall roughly into two categories: (*a*) those based on standard bonding and van der Waals interactions that are used in conventional molecular dynamics studies and (*b*) knowledge-based potentials that are derived from known structures. The functions based on quantum mechanical calculations and fit to spectroscopic data on small molecules have been well parameterized to describe the motion of the atoms about or near their crystallographic coordinates. It is not yet possible to test whether they are sufficient to describe the folding process, since the natural process is very slow on the time scale of atomistic simulations ($\approx$1 ns). Thus the theoretician must seek simpler interaction functions that can encode sequence-structure relationships. There is much raw material for trying to learn these relationships. According to the latest release of the SWISSPROT database, we know the sequences $\{S_i\}$ of about 40,000 proteins. A small subset of these proteins, approximately 4000, have been well crystallized or are small enough for structure determination by NMR, so the mean positions $\{r_i\}$ of all the atoms have been determined. Energy landscape theory provides a way of understanding the learning process of extracting important correlations from these raw data. Indeed the energy landscape theory of structure prediction has many mathematical parallels with the theories of learning used in connectionist artificial intelligence (204, 205). Using the fact that evolution has already found out how to design different sequences $\{S_i\}$ compatible with a given structure $\{r_i\}$, scientists trying to break the protein folding code must use all the information collected in the sequence and structural databases along with such an organized theory of inference.

While the landscape energy ideas provide a way for understanding the algorithms, the idea of extracting energy parameters from a statistical analysis of the

structural database predates it and originated with the early papers of Tanaka & Scheraga (206) and Miyazawa & Jernigan (135, 136). In a similar approach, Sippl and coworkers (207, 208) calculated distance-dependent potentials that included information about the proximity of the residues in the sequence. Eisenberg and coworkers (209, 210) developed effective one-body profile potentials for each amino acid from studying its context in typical structures. None of these early methods made explicit use of the energy landscape theory. Therefore it is no surprise that in application they do not satisfy the minimal frustration criterion and that multiple minima often plague their use. Since the criterion that the native state must have a pronounced energy minimum with respect to the distribution of misfolded conformations is so important for optimization algorithms to successfully find the global minimum, it is interesting to use it directly in the design of energy functions.

An approach that puts the learning problem in direct contact with energy landscape ideas is the use of the associative memory Hamiltonian (AMH). In 1989 Friedrichs & Wolynes introduced an AMH ($\mathcal{H}_{\mathrm{AMH}}$) that encodes correlations between the sequence of the target protein whose structure is to be determined and the sequences and structures of a representative set of memory proteins taken from a database (102). In one respect the AMH resembles the empirical energy functions used in conventional molecular dynamics, by describing an effective backbone potential $\mathcal{V}_o$ common to all proteins. Such a chain molecule potential for the backbone atoms includes harmonic terms to induce backbone rigidity and the correct chirality, and to prevent the overlap of nonbonded atoms. The correlations in the database are used directly to find the sequence-dependent interaction energies. This is analogous to the way in which neural networks are built to perform pattern recognition. The novel aspect of the AMH is its use of pair interactions defined by analogy to the connections found in pattern recall neural networks. Since the AMH is actually used to generalize and generate structures for new sequences and not just the old ones, the name is perhaps not the best—they really are pattern recognition Hamiltonians. The Hamiltonian has the form (51, 155, 156)

$$\mathcal{H}_{\mathrm{AMH}}(\{r_{ij}, A_i\}) = -\sum_{\mu}\sum_{i<j}\gamma_{ij}^{\mu}\theta\left(r_{ij} - r_{i'j'}^{\mu}\right) + \mathcal{V}_o. \qquad 22.$$

The first term, the associative memory potential, is a function of the pairwise distance between selected atoms (usually $C-\alpha$ or $C-\beta$) of residues i and j, r_{ij}. γ_{ij}^{μ} encodes a degree of similarity between residues i and j of the target protein and a corresponding pair in the memory protein μ, and it may include information about the physicochemical properties of the residues, their probability of mutation, or context of the residues in the protein. $\theta(r_{ij} - r_{i'j'}^{\mu})$ is chosen for convenience to be a Gaussian function of the distance of the

difference between the pairwise distance in the target structure and the memory structure. Notice that if the database is extremely large and we use no particular rule for associating pairs of residues in the memory with those in the target, the energy function would be a simple statistical pair correlation function of the properties encoded by γ in the known database. In that sense the AMH is just a more general form of pair potential than the statistical pair correlations introduced above. On the other hand, for a finite database and with limited rules of correspondence, the AMH can be analyzed in accordance with energy landscape theory very much in the way Hopfield neural networks have been studied using spin glass theory. We can see that for a small database, the energy function has many minima of varying depth corresponding with the structure of each database or memory protein as well as mixed minima. These minima are differentiated by the sequence property similarity weights γ^{μ}.

In essence, each memory protein constructs a small folding funnel. If these funnels add coherently or only a single one dominates, the Hamiltonian will not be very frustrated, and a single folding funnel to a structure consistent with empirical correlations will be formed. On the other hand, if the database examples conflict with the properties chosen to measure sequence similarity a rugged landscape results and optimization will give different results dependent on the starting point. Friedrichs, Wolynes, and coworkers showed that with naive encodings, the size database that could be accommodated was large but still inadequate for useful predictions (156). Achieving a coherent addition of the funnels, even for large databases, that allows generalization to novel structures requires optimization of the γ-parameters for a given encoding and selection of appropriate training proteins with their corresponding memory proteins. This is where the quantitative version of the principle of minimum frustration was used. Maximizing the dimensionless ratio T_F/T_G for a set of training proteins produces a problem very much like the variational principle used in quantum chemistry. Goldstein et al (51) showed that with the simple energy landscape analysis the variationally optimal energy parameters can be expressed in terms of two statistical quantities giving the gap and the ruggedness of the landscapes for the training examples. They write the average gap $\delta E_s = \mathbf{A}\gamma$ and the average ruggedness $\Delta E^2 = \gamma \mathbf{B} \gamma$, where $\mathbf{A}$ is a vector and $\mathbf{B}$ a matrix given by the statistical quantities

$$\mathbf{A}_i = \langle \lambda_i \rangle_f - \langle \zeta_i \rangle_{mis} \qquad 23.$$

and

$$\mathbf{B}_{ij} = \langle \zeta_i \zeta_j \rangle_{mis} - \langle \zeta_i \rangle_{mis} \langle \zeta_j \rangle_{mis}. \qquad 24.$$

λ_i denotes the frequency of γ_i interaction in the folded structures, and ζ_i denotes the frequency of γ_i in the full ensemble of misfolded structures. The explicit

values of **A** and **B** are obtained by averaging over a set of training proteins with known structures and simulated molten globule states. According to Equation 12, T_F/T_G is maximized when the dimensionless ratio $R = \delta E_s/\sqrt{\Delta E^2}$ is maximized. Solving this variational problem is straightforward and does not have multiple minima. This relative lack of frustration for the decoding task is important and also relevant for protein design. In fact, this maximization procedure leads to an explicit formula $\gamma = \mathbf{B}^{-1}\mathbf{A}$. Simulations with naive assignments of γ, e.g. interaction energies between similar residues in the target and memory being $E_{sim} = -3$ and between dissimilar residues $E_{dis} = -1$, give rise to much smaller T_F/T_G values than with optimization. For an optimized comparison code based only on hydrophobicity and proximity, the results of molecular dynamics runs are much more encouraging. Some snapshots of runs for the small helical protein 2cro using a single correct memory are shown in Figure 15. The simulation begins with the protein in a random extended form and shows the collapse and compaction of the protein to its native fold.

This same optimization strategy for learning potentials can be used to learn parameters in the more conventional forms for potentials that do not explicitly

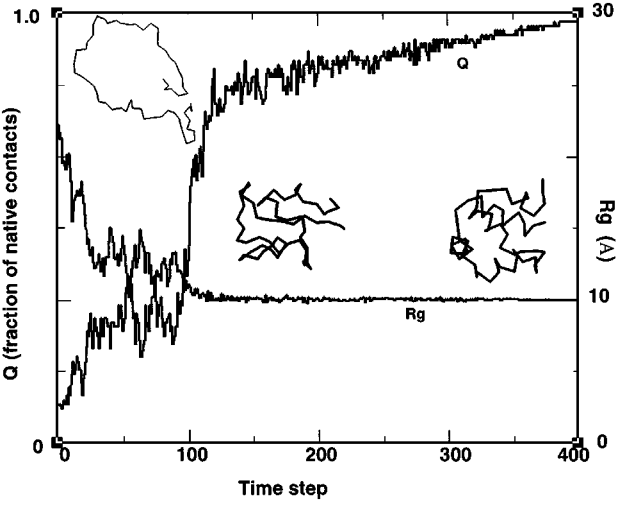

Figure 15 Simulated annealing run showing the collapse and structure formation of the small helical protein 2cro ($N = 65$) using the $\mathcal{H}_{AMH}$ energy function with one memory. The simulation begins with the protein in a random extended form with a radius of gyration typical of a random coil, $R_g = 30$ Å. Collapse and compaction of the protein occurs quickly to a state that has roughly the native topology or fold but has incomplete secondary structure. Continued folding in this compact state completes the formation of the helices and modifies the tertiary contacts.

match target proteins with memories (80). These are of course less discriminating because their T_F/T_G ratios are smaller. But they can still be used for screening predictions of protein structures in which different candidates are compared. An interesting model study of the optimization strategy for deriving potentials has been carried out by Mirny & Shakhnovich (211). Using a somewhat different averaging scheme from Goldstein and collaborators (51, 80), Mirny & Shakhnovich (211) also infer a contact Hamiltonian from a set of lattice proteins designed originally with the minimal frustration principle. They show that the use of the optimization learning scheme gives back a potential that reproduces the structures. While the exact method in which training proteins are averaged over might seem to make a significant difference, their study shows only a few percent improvement in the discrimination score over the averaging scheme presented by Goldstein et al (51, 80).

What is the relation between the optimization learning schemes and the older approaches to learning pair potentials? The older approach to extracting structure-based potentials assumes there is a potential of mean force for the experimentally observed frequencies of nonbonded amino acid residue pair contacts. This assumption is puzzling until examined with energy landscape theory. The potential of mean force W between two residues in contact at a distance r_c is assumed to be related to these database frequencies through a Boltzmann factor

$$\exp\{-W_{ij}(r_c)/RT\} = \frac{P_{ij}(r_c)}{P_{ij}(ref)}, \qquad 25.$$

where $P_{ij}(ref)$ is a reference probability and T is some effective temperature usually left undetermined. Using the quasi-chemical approximation that neglects chain connectivity, Miyazawa, Jernigan, and coworkers (136, 212, 213) described the reference state in terms of various random mixture approximations. This gives rises to an effective contact energy E_{ij} that is measured relative to either self-interactions

$$E_{ij} = W_{ij}(r_c) - [W_{ii}(r_c) + W_{jj}(r_c)]/2 \qquad 26.$$

or interactions with some average residue or solvent molecule A

$$E_{ij} = W_{ij}(r_c) + W_{AA}(r_c) - W_{iA}(r_c) - W_{jA}(r_c). \qquad 27.$$

In the simplest application of these potentials the interaction sites are typically chosen to be the position of the C_α atom or some atom in the side chain of each amino acid residue. Sippl (208) and others have also constructed potentials of mean force using radial distribution functions $g_{ij}(r, l)$ for pairs of residues separated by a distance of l in the sequence

$$W_{ij}(r, l) = -RT \log g_{ij}(r, l). \qquad 28.$$

Energy landscape theory provides an interpretation of the effective temperatures in these expressions, if we neglect the principle of minimal frustration. If the REM without minimal frustration is assumed valid for natural proteins (i.e. natural proteins are truly random heteropolymers), the spin glass analogy used in energy landscape theory suggests that the probability of a particular ground state energy is given by a Boltzmann distribution at the glass transition temperature (111, 214). Thus the number of states near E_g is approximately,

$$n(E) = \exp\left(\frac{E - E_{\rm g}}{T_{\rm G}}\right), \qquad 29.$$

and if the energy E of a conformation is assumed to be a sum of independent pairwise contact energies, then the frequencies f_{ij} of contacts would be $f_{ij} = \exp(-\Delta W_{ij}/T_{\rm G})$. Here the energies are measured from the average energy per contact at $T_{\rm G}$. Again assuming the quasi-chemical approximation to be valid, the log probability formula is obtained and the mysterious effective temperature can be identified with the $T_{\rm G}$.

The knowledge-based potentials of mean force have often been successfully used to evaluate the compatibility of a sequence with a given structure. The contact energies generated either from the frequencies or the radial distribution functions have been used primarily to evaluate the compatibility of a sequence on a given structure as in the work of Maiorov & Crippen (215). Some of the progress in this area has been reviewed by Wodak and coworkers (216, 217). The successful implementation of these Boltzmann-weighted potentials into prediction by full folding simulations has not been demonstrated. Indeed, recent lattice simulations by Hinds & Levitt (218) using contact energies based on the procedure of Miyazawa & Jernigan (212) have generated native-like conformations with about 30% of the native contacts in place. This is consistent with the fact that the principle of minimal frustration was explicitly not used in deriving the potentials. The difference between these two types of approaches for finding database potentials depends on how truly optimized for folding natural proteins are. Determination of $T_{\rm F}/T_{\rm G}$ by experiments is central to providing an answer. Whether any scheme will work may depend on the level of description used. More encouraging results have been reported when a much finer lattice grid is incorporated with a more detailed potential that includes many-body terms and orientation factors (219, 220).

Seemingly different optimization strategies not based on Monte Carlo or molecular dynamics used to predict protein structures can also be understood with energy landscape theory ideas. An important practical procedure is the threading algorithm in which one tries to thread a new sequence onto all the known structures. It is based on the idea that at least the more common protein structures are limited in number. The distribution of natural folds is interesting.

Jones & Thornton have shown that certain super folds dominate the current database (221). Goldstein and coworkers have explained this using energy landscape theory (222), and it has also been shown in an interesting lattice study by Tang and coworkers (223, 224). Indeed, the total number of structures, may be limited. Of the current 4000 structures, there are 200 different fold topologies, and Chothia has estimated that this number will increase to only about 1000 when the sequencing of the human genome is completed (225, 226). Threading tries to match structure with sequence in a fashion similar to the way sequences are matched onto each other in phylogenetic analysis. Pioneering papers showing this possibility came from the groups of Eisenberg and coworkers (209) and Thornton and coworkers (227), and since them many other similar schemes have been used with potentials of mean force learning of the energy function. In the most general scheme of this type, the evaluation of this sequence-structure alignment is based on contributions from pair contact terms, $E_{\rm ct}$, a pseudo one-body profile ($E_{\rm p}$), hydrogen bonding ($E_{\rm hb}$), gaps in the aligned structure ($E_{\rm g}$), and satisfaction of any known experimental constraints (80, 214, 228, 229):

$$E_{\rm T} = E_{\rm p} + E_{\rm ct} + E_{\rm hb} + E_{\rm g} + E_{\rm exp}. \qquad 30.$$

The profile energy ($E_{\rm p}$) is a measure of the propensity of an amino acid A_i to reside in a particular context of secondary structure SS_i and surface accessibility SA_i.

$$E_{\rm p} = \sum_{i=1}^{N} \gamma^{p}(A_i, SS_i, SA_i) \qquad 31.$$

The contact energy $E_{\rm ct}$ measures the pairwise interaction energies within two cut-off radii and monitors selected multibody interactions (*mb*) such as multiple cysteine bond formations:

$$E_{\rm ct} = \sum_{i=1}^{N-2} \sum_{j=i+2}^{N} \sum_{k=1}^{2} \gamma_k^{ct}(A_i, A_j) u\left(r_k^{\rm ct} - r_{ij}\right) + mb. \qquad 32.$$

r_{ij} is the distance between residues A_i and A_j, and in simple models $u(r_{1,2}^{\rm ct} - r_{ij})$ can be taken as unit step functions that include all short range interactions acting within 5 Å and all intermediate range interactions between 5 and 12 Å, respectively. The energy term $E_{\rm hb}$ provides contributions for backbone hydrogen bond formation within α-helices and between β-sheets. The gap energy term ($E_{\rm g}$) enforces only physically acceptable insertions, deletions, or bulges in the sequence-structure alignment. The energy term ($E_{\rm exp}$) may be used to aid in guiding an alignment to incorporate experimental data such as known contacts in an active site. The explicit use of energy landscape theory improves

the performance of threading schemes. In early threading algorithms the gap parameters are usually inferred empirically, but the energy landscape theory allows them to be chosen in an optimal way again by optimizing a stability gap to ruggedness ratio (214, 228). The simple energy landscape learning approaches described above can be improved upon when correlations in the energy landscape are taken into account (230). Rather than using the REM approximation for the competing alternative states, a self-consistent optimization can be carried out to improve the discrimination. In this scheme, shown in Figure 16, low-energy misfolded structures are explicitly constructed in computing the gap and ruggedness. This has so far been only carried out successfully for energy functions used in threading because low-energy minima are easy to generate by this method (229). The self-consistent energy function produces low rms threading alignments for distant homologs and was used in Figure 17 to predict a structure for the archeael AK sequence shown in Figure 1 (231).

Many of the same landscape ideas used earlier to develop protein structure prediction methods have their counterpart in protein design. We may say that protein structure prediction is reverse engineering, i.e. trying to tease out the rules nature has used in design. If we are bold enough to assume those rules already known, then the same mathematical principles can be used for forward engineering, i.e. sequence design for a given structure. So far this has only been tested in the artificial world of lattice models, but the study has been enlightening. Shakhnovich and coworkers used the simple form of the minimal frustration principle to design sequences of lattice proteins with simple codes (232–234). For the two letter codes the preliminary results were very encouraging, but in a contest between Dill, Shakhnovich, and coworkers the simple scheme was found to fail (235). The origin of the failure for the two letter code lies in the lack of self-consistency in the optimization, since the MG states were in fact microphase separated, i.e. correlations exist in the landscape. The situation for the many-letter code even with the simple approximation seems much better. Hao & Scheraga (176) and Deutsch & Kurosky (236) have recently proposed a design scheme using self-consistent optimization. This performs considerably better. Clearly, to be of practical significance design schemes must be developed for use with reliable atomistic potentials or several rounds of engineering and reverse engineering (i.e. learning) will have to be carried out. This is likely to be a major activity of the future.

The theory of protein folding based on energy landscapes provides a general framework for thinking about many facets of this complex problem. The next stage involves many developments. Clearly, the formal theory of folding using energy landscapes is still developing. At the same time experimental techniques that allow the submillisecond stages of folding are becoming available. Protein engineering is allowing site-specific experimental probes to be introduced, and

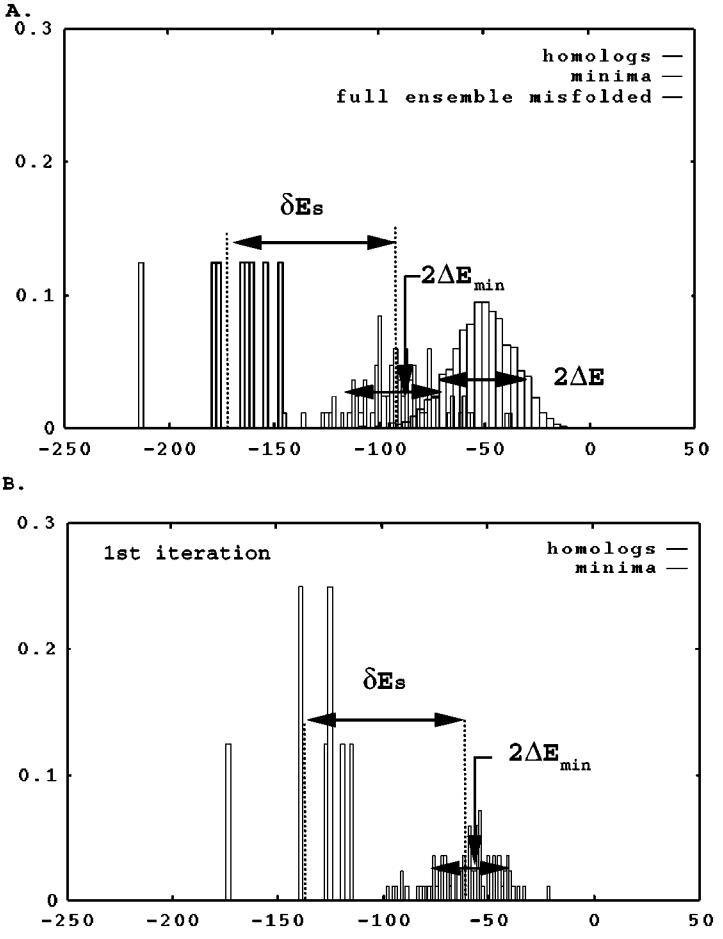

Figure 16 The distributions of energy states for a typical training protein before and after one round of self-consistency. (*A*) Using the original optimized energy function, the distribution of the thermal minima of the misfolded states (Gaussian-like curve at the left) lies close to the energies of the native and homolog structures. (*B*) By re-optimizing with respect to the mean of these minima, the energy function is able to better discriminate the correct (native and homologs) structures from the minima that have many features of real proteins such as partial redistribution of hydrophobic residues toward the interior of the protein due to microphase separation (229).

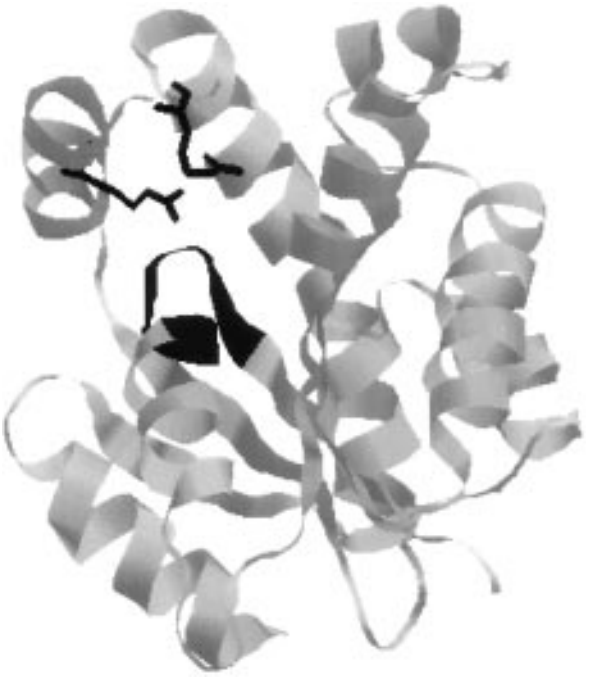

Figure 17 Predicted structure of adenylate kinase from *Methanococcus jannaschii*. The structure was obtained by a mean-field alignment of the archaeal sequence given in Figure 1 to the scaffold of an uridylate kinase from yeast. The sequence-structure compatibility during the alignment was evaluated according to Equation 30. Residues believed to be essential in its function to catalyze the biosynthesis of ADP from ATP and AMP are shown in black.

new information about the correlations of structures in partially folded protein ensembles is becoming available. The practical aspects of protein science are also acting to encourage the development of the theoretical perspective. Structure prediction and actual design of proteins foldable in the laboratory look to be in sight. The current theory provides a route to achieving these goals.

ACKNOWLEDGMENTS

We would like to thank W Eaton and J Bryngelson for reading the manuscript. We also thank N Socci, I Balabin, H Nymeyer, Kris Koretke, and Ben Shoemaker for help with the preparation of some figures. Work at UCSD was supported by the National Science Foundation (Grant no. MCB–9603839) and the UC/Los Alamos Research (CULAR) Initiative, and at UIUC by the National Institutes of Health (Grant no. 1R01 GM44557). This paper was finished while PGW was

a Fogarty Scholar in residence at the National Institutes of Health at Bethesda, MD.

Literature Cited

1. Watson JD, Crick FHC. 1953. *Nature* 177:964
2. Pauling L, Corey RB, Branson HR. 1951. *Proc. Natl. Acad. Sci. USA* 37:205–11
3. Anfinsen C, Haber E, Sela M, White F Jr. 1961. *Proc. Natl. Acad. Sci. USA* 47:1309
4. Anfinsen CB. 1973. *Science* 181:223–30
5. Kauzmann W. 1959. *Adv. Protein Chem.* 16:1–64
6. Poland D, Scheraga HA. 1970. *Theory Of Helix-Coil Transitions In Biopolymers.* New York: Academic
7. Englander SW, Mayne L. 1992. *Annu. Rev. Biophys. Biomol. Struct.* 21:243–65
8. Creighton TE. 1990. *Biochem. J.* 270:1–16
9. Kim PS, Baldwin RL. 1990. *Annu. Rev. Biochem.* 59:631–60
10. Bryngelson JD, Wolynes PG. 1987. *Proc. Natl. Acad. Sci. USA* 84:7524–28
11. Bryngelson JD, Wolynes PG. 1989. *J. Phys. Chem.* 93:6902–15
12. Garel T, Orland H, Thirumalai D. 1996. In *New Developments in Theoretical Studies of Proteins*, ed. R Elber, pp. 197–268. Singapore: World Sci.
13. Wolynes PG. 1991. In *Carges Lectures 1990 in Biologically Inspired Physics*, ed. L Peliti, pp. 15–37. New York: Plenum
14. Wolynes PG. 1992. In *Spin Glasses and Biology*, ed. D Stein, pp. 225–59. Singapore: World Sci.
15. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. 1995. *Proteins* 21:167–95
16. Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, et al. 1995. *Protein Sci.* 4:561–602
17. Wolynes PG, Onuchic JN, Thirumalai D. 1995. *Science* 267:1619–20
18. Mirny LA, Abkevich V, Shakhnovich EI. 1996. *Folding and Design* 1:103–16
19. Fersht AR. 1997. *Curr. Opin. Struct. Biol.* 7:3–9
20. Eaton WA, Munoz V, Thompson PA, Chan CK, Hofrichter J. 1997. *Curr. Opin. Struct. Biol.* 7:10–14
21. Eaton WA, Thompson P, Chan CK, Hagen SJ, Hofrichter J. 1996. *Structure* 4:1133–39
22. Leopold PE, Montal M, Onuchic JN. 1992. *Proc. Natl. Acad. Sci. USA* 89: 8721–25
23. Onuchic JN, Wolynes PG, Luthey-Schulten Z, Socci ND. 1995. *Proc. Natl. Acad. Sci. USA* 92:3626–30
24. Huang GS, Oas TG. 1995. *Proc. Natl. Acad. Sci. USA* 92:6878–82
25. Feng Y, Sligar SG, Wand AJ. 1994. *Nat. Struct. Biol.* 1:30–36
26. Nash D, Lee B, Jonas J. 1996. *Biochim. Biophys. Acta* 1297:40–48
27. Roder H. 1995. *Nat. Struct. Biol.* 2:817–20
28. Elove GA, Bhuyan AK, Roder H. 1994. *Biochemistry* 33:6925–35
29. Wright PE, Jennings PA. 1993. *Science* 262:892–95
30. Dyson HJ, Wright PE. 1996. *Annu. Rev. Phys. Chem.* 47:369–95
31. Burton R, Huang GS, Daugherty MA, Fullbright PW, Oas TG. 1996. *J. Mol. Biol.* 263:311–22
32. Alexandrescu A, Evans P, Pitkeathly M, Baum J, Dobson C. 1993. *Biochemistry* 32:1707–18
33. Balbach J, Forge V, Lau WS, Vannuland N, Brew K, Dobson CM. 1996. *Science* 274:1161–63
34. Plaxco KW, Dobson CM. 1996. *Curr. Opin. Struc. Biol.* 6:630–36
35. Itzhaki LS, Otzen DE, Fersht AR. 1995. *J. Mol. Biol.* 254:260–88
36. López-Hernández E, Serrano L. 1996. *Folding Design* 1:43–55
37. Jones CM, Henry ER, Hu Y, Chan C-K, Luck SD, et al. 1993. *Proc. Natl. Acad. Sci. USA* 90:11860–64
38. Phillips CM, Mizutani Y, Hochstrasser RM. 1995. *Proc. Natl. Acad. Sci. USA* 92:7292–96
39. Williams S, Causgrove TP, Gilmanshin R, Fang KS, Callender RH, et al. 1996. *Biochemistry* 35:691–97
40. Pascher T, Chesick JP, Winkler JR, Gray HB. 1996. *Science* 271:1558–60
41. Ballew RM, Sabelko J, Gruebele M. 1996. *Proc. Natl. Acad. Sci. USA* 93:5759–64

42. Ballew RM, Sabelko J, Gruebele M. 1996. *Nat. Struct. Biol.* 3:923–26
43. Chan C-K, Hu Y, Takahashi S, Rousseau DL, Eaton WA, Hofrichter J. 1997. *Proc. Natl. Acad. Sci. USA* 94:1779–84
44. Derrida B. 1980. *Phys. Rev. Lett.* 45:79–82
45. Derrida B. 1981. *Phys. Rev. B* 24:2613–26
46. Frauenfelder H, Parak F, Young RD. 1988. *Annu. Rev. Biophys. Biophys. Chem.* 17:451–79
47. Frauenfelder H, Sligar SG, Wolynes PG. 1991. *Science* 254:1598–603
48. Stein DL. 1985. *Proc. Natl. Acad. Sci. USA* 82:3670–72
49. Frauenfelder H, Alberding NA, Ansari A, Braunstein D, et al. 1990. *J. Phys. Chem.* 94:1024–37
50. Flory PJ. 1956. *J. Am. Chem. Soc.* 78:5222–35
51. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. 1992. *Proc. Natl. Acad. Sci. USA* 89:4918–22
52. Dill KA. 1990. *Biochemistry* 29:7133–55
53. Dinner A, Šali A, Karplus M, Shakhnovich E. 1994. *J. Chem. Phys.* 101:1444–51
54. Socci ND, Onuchic JN. 1995. *J. Chem. Phys.* 103:4732–44
55. Iben I, Braunstein D, Doster W, Frauenfelder H, Hong M, et al. 1989. *Phys. Rev. Lett.* 62:1916–19
56. Tilton R, Dewan J, Petsko G. 1992. *Biochemistry* 31:2469–81
57. Ansari A, Jones CM, Henry ER, Hofrichter J, Eaton WA. 1992. *Science* 256:1796–98
58. Sochava IV, Smirnova OI. 1993. *Food Hydrocolloids* 6:513–24
59. Angell CA. 1995. *Science* 267:1924–35
60. Wolynes PG, Schulten ZL, Onuchic J. 1996. *Chemistry* and *Biology* 3:425–32
61. Chan HS, Dill KA. 1991. *Annu. Rev. Biophys. Biophys. Chem.* 20:447–90
62. Bryngelson JD, Wolynes PG. 1990. *Biopolymers* 30:177–88
63. Dyson HJ, Wright PE. 1991. *Annu. Rev. Biophys. Biophys. Chem.* 20:519–38
64. Janin J, Wodak S. 1983. *Prog. Biophys. Mol. Biol.* 42:21
65. Honig B, Cohen FE. 1996. *Folding Design* 1:R17–R20
66. Hamada D, Segawa S, Goto Y. 1996. *Nat. Struct. Biol.* 3:868–73
67. Luthey-Schulten Z, Ramirez BE, Wolynes PG. 1995. *J. Phys. Chem.* 99: 2177–85
68. Flory PJ. 1956. *Proc. R. Soc. London Ser. A* 234:73–89
69. Flory PJ. 1969. *Statistical Mechanics of Chain Molecules.* New York: Wiley Interscience.
70. Matheson RR, Flory PJ. 1981. *Macromolecules* 14:954–60
71. Onsager L. 1949. *Ann. NY Acad. Sci.* 51:627–59
72. Grover MK, Zwanzig R. 1974. *Biopolymers* 13:2103–15
73. Bascle J, Garel T, Orland H. 1993. *J. Physique II* 3:245–53
74. Saven JG, Wolynes PG. 1996. *J. Mol. Biol.* 257:199–216
75. Derrida B. 1985. *J. Physique Lett.* 46: L401–7
76. Derrida B, Gardner E. 1986. *J. Phys. C* 19:2253–74
77. Plotkin SS, Wang J, Wolynes PG. 1996. *Phys. Rev. E* 53:6271–96
78. Bryngelson JD, Thirumalai D. 1996. *Phys. Rev. Lett.* 76:542–45
79. Gutin A, Shakhnovich E. 1994. *J. Chem. Phys.* 100:5290–93
80. Goldstein R, Luthey-Schulten ZA, Wolynes PG. 1992. *Proc. Natl. Acad. Sci. USA* 89:9029–33
81. Pande VS, Grosberg AY, Joerg C, Tanaka T. 1996. *Phys. Rev. Lett.* 76:3987–90
82. Plotkin SS, Wang J, Wolynes PG. 1997. *J. Chem. Phys.* 106:2932–48
83. Mezard M, Parisi G, Virasoro MA. 1986. *Spin Glass Theory and Beyond.* Singapore: World Sci.
84. Rammal R, Toulouse G, Virasoro MA. 1986. *Rev. Mod. Phys.* 58:765–88
85. Edwards SF, Muthukumar M. 1988. *J. Chem. Phys.* 89:2435–41
86. Obukhov SP. 1990. *Phys. Rev. A* 42:2015–19
87. Honeycutt JD, Thirumalai D. 1989. *J. Chem. Phys.* 90:4542–59
88. Fernández A. 1989. *J. Phys. A* 22:3137–42
89. Garel T, Orland H. 1988. *Europhys. Lett.* 6:597–601
90. Garel J, Garel T, Orland H. 1989. *J. Physique* 50:3067–74
91. Shakhnovich EI, Gutin AM. 1989. *Biophys. Chem.* 34:187–99
92. Edwards SF, Anderson PW. 1975. *J. Phys. F* 5:965–74
93. Garel T, Orland H. 1988. *Europhys. Lett.* 6:307–10
94. Gross DJ, Kanter I, Sompolinsky H. 1985. *Phys. Rev. Lett.* 55:304–97
95. Kirkpatrick TR, Wolynes PG. 1987. *Phys. Rev. B* 36:8552–64
96. Gutin AM, Shakhnovich EI. 1993. *J. Chem. Phys.* 98:8174–77
97. Sfatos CD, Gutin AM, Shakhnovich EI. 1993. *Phys. Rev. E* 48:465–75

98. Sfatos CD, Gutin AM, Shakhnovich EI. 1994. *Phys. Rev. E* 50:2898–905
99. Pande VS, Grosberg AY, Tanaka T. 1995. *Phys. Rev. E* 51:3381–92
100. Sasai M, Wolynes PG. 1990. *Phys. Rev. Lett.* 65:2740–43
101. Sasai M, Wolynes PG. 1992. *Phys Rev A* 46:7979–97
102. Friedrichs MS, Wolynes PG. 1989. *Science* 246:371–73
103. Mezard M, Parisi G. 1991. *J. Physique I* 1:809–36
104. Takada S, Wolynes PG. 1997. *Phys. Rev. E* 55:4562–77
105. LaBean TH, Kauffman SA, Butt TR. 1995. *Mol. Diversity* 1:29–38
106. Davidson AR, Sauer RT. 1994. *Proc. Natl. Acad. Sci. USA* 91:2146–50
107. Ramanathan S, Shakhnovich E. 1994. *Phys. Rev. E* 50:1303–12
108. Pande VS, Grosberg AY, Tanaka T. 1995. *Macromolecules* 28:2218–27
109. Cairns-Smith AG. 1990. *Seven Clues to the Origin of Life: A Scientific Detective Story.* Cambridge: Cambridge Univ. 2nd ed.
110. Finkelstein AV, Gutin AM, Badretdinov AY. 1993. *FEBS Lett.* 325:23–28
111. Finkelstein AV, Gutin AM, Badretdinov AY. 1995. *Proteins: Struct. Funct. Genet.* 23:142–49
112. Wolynes PG. 1996. *Proc. Natl. Acad. Sci. USA* 93:14249–55
113. Kellman ME. 1996. *J. Chem. Phys.* 105:2500–8
114. Nelson ED, Eyck LT, Onuchic JN. 1997. *Phys. Rev. Lett.* submitted
115. Kirkpatrick TR, Thirumalai D, Wolynes PG. 1989. *Phys. Rev. A* 40:1045–54
116. Fisher DS, Huse DA. 1988. *Phys. Rev. B* 38:386–411
117. Fisher DS, Huse DA. 1988. *Phys. Rev. B* 38:373–85
118. McMillan WL. 1985. *Phys. Rev. B* 31: 340–41
119. Thirumalai D. 1995. *J. Physique I* 5: 1457–67
120. Franz S, Mézard M, Parisi G. 1993. *Int. J. Neural Syst.* 3:195–200
121. Baldwin RL. 1996. *Folding Design* 1:R1–R8
122. Brooks CL, Case D. 1993. *Chem. Rev.* 93:2487–502
123. Simmerling C, Elber R. 1994. *J. Am. Chem. Soc.* 116:2534–47
124. Hirst JD, Brooks CL. 1994. *J. Mol. Biol.* 243:173–78
125. Hirst JD, Brooks CL. 1995. *Biochemistry* 34:7614–21
126. Boczko EM, Brooks CL. 1995. *Science* 269:393–96
127. Daggett V, Levitt M. 1992. *Proc. Natl. Acad. Sci. USA* 89:5142–46
128. Daggett V, Levitt M. 1993. *J. Mol. Biol.* 232:600–19
129. Daggett V, Li A, Itzhaki LS, Otzen DE, Fersht AR. 1996. *J. Mol. Biol.* 257:430–40
130. Hünenberger PH, Mark AE, van Gunsteren WF. 1995. *Proteins* 21:196–213
131. Mark AE, van Gunsteren WF. 1992. *Biochemistry* 31:7745–48
132. Garcia AE, Hummer G. 1996. *Personal communication.*
133. Scheraga HA. 1971. *Chem. Rev.* 71:195–217
134. Gō N. 1983. *Annu. Rev. Biophys. Bioeng.* 12:183–210
135. Miyazawa S, Jernigan RL. 1982. *Biopolymers* 21:1333
136. Miyazawa S, Jernigan RL. 1985. *Macromolecules* 218:534–52
137. Covell DG, Jernigan RL. 1990. *Biochemistry* 29:3287–94
138. Covell DG. 1994. *J. Mol. Biol.* 235:1032–43
139. Lau KF, Dill KA. 1989. *Macromolecules* 22:3986
140. Chan HS, Dill KA. 1989. *Macromolecules* 22:4559–73
141. Shakhnovich E, Farztdinov G, Gutin AM, Karplus M. 1991. *Phys. Rev. Lett.* 67:1665–68
142. Shakhnovich EI, Gutin AM. 1993. *Proc. Natl. Acad. Sci. USA* 90:7195
143. Abkevich VI, Gutin AM, Shakhnovich EI. 1994. *Biochemistry* 33:10026–36
144. Šali A, Shakhnovich E, Karplus M. 1994. *J. Mol. Biol.* 235:1614–36
145. Camacho CJ, Thirumalai D. 1993. *Phys. Rev. Lett.* 71:2505–8
146. Hao M-H, Scheraga HA. 1994. *J. Phys. Chem.* 98:4940–48
147. Pande VS, Grosberg AY, Tanaka T. 1994. *Proc. Natl. Acad. Sci. USA* 91:12976–79
148. Socci ND, Onuchic JN. 1994. *J. Chem. Phys.* 101:1519–28
149. Skolnick J, Kolinski A. 1990. *J. Mol. Biol.* 212:787–817
150. Skolnick J, Kolinski A. 1990. *Science* 250:1121–25
151. Kolinski A, Milik M, Skolnick J. 1991. *J. Chem. Phys.* 94:3978–85
152. Godzik A, Kolinski A, Skolnick J. 1993. *J. Comp. Chem.* 14:1194–2
153. Skolnick J, Galazka W, Kolinski A. 1995. *J. Chem. Phys.* 103:10286–97
154. Levitt M, Warshel A. 1975. *Nature* 253:694–98
155. Friedrichs M, Wolynes PG. 1990. *Tet. Comp. Meth.* 3:175

156. Friedrichs MS, Goldstein RA, Wolynes PG. 1991. *J. Mol. Biol.* 222:1013–34
157. Guo Z, Thirumalai D, Honeycutt JD. 1992. *J. Chem. Phys.* 97:525–35
158. Honeycutt JD, Thirumalai D. 1990. *Proc. Natl. Acad. Sci. USA* 87:3526–29
159. Honeycutt J, Thirumalai D. 1992. *Biopolymers* 32:695–709
160. Thirumalai D, Guo ZY. 1995. *Biopolymers* 35:137–40
161. Ball KD, Berry RS, Kunz R, Li F-Y, Proykova A, Wales DJ. 1996. *Science* 271:963–66
162. Wales DJ. 1996. *Science* 271:925–29
163. Doye JPK, Wales DJ. 1996. *J. Chem. Phys.* 105:8428–45
164. Rose JP, Berry RS. 1996. Poster Abstr. Los Alamos
165. Sasai M. 1995. *Proc. Natl. Acad. Sci. USA* 92:8438–42
166. Monge A, Friesner RA, Honig B. 1994. *Proc. Natl. Acad. Sci. USA* 91:5027–29
167. Friesner RA, Gunn JR. 1996. *Annu. Rev. Biophys. Biomol. Struct.* 25:315–42
168. Verdier PH, Stockmayer WH. 1962. *J. Chem. Phys.* 36:227–35
169. Hilhorst HJ, Deutch JM. 1975. *J. Chem. Phys.* 63:5153–61
170. Kremer K, Binder K. 1988. *Comp. Phys. Rept.* 7:259–310
171. Gurler MT, Crabb CC, Dahlin DM, Kovac J. 1983. *Macromolecules* 16:398–403
172. Shakhnovich E, Abkevich V, Ptitsyn O. 1996. *Nature* 379:96–98
173. Skolnick J, Kolinski A. 1991. *J. Mol. Biol.* 221:499–531
174. Socci ND, Onuchic JN, Wolynes PG. 1996. *J. Chem. Phys.* 104:5860–68
175. Hao M-H, Scheraga HA. 1995. *J. Chem. Phys.* 102:1334–48
176. Hao M, Scheraga H. 1996. *Proc. Natl. Acad. Sci. USA* 93:4984–89
177. Klimov DK, Thirumalai D. 1996. *Proteins: Struct. Funct. Genet.* 26:411–41
178. Shrivastava I, Vishveshwara S, Cieplak M, Maritan A, Banavar JR. 1995. *Proc. Natl. Acad. Sci. USA* 92:9206–9
179. Creighton TE. 1993. *Proteins: Structure and Molecular Properties.* New York: Freeman 2nd ed.
180. Onuchic JN, Socci ND, Luthey-Schulten Z, Wolynes PG. 1996. *Folding Design* 1:441–50
181. Frenkel J. 1946. *Kinetic Theory of Liquids.* London: Oxford Claredon
182. Ma S-K. 1985. *Statistical Mechanics.* Philadelphia: World Sci.
183. Finkelstein AV, Badretdinov AY. 1997. *Folding Design* 2:115–21
184. Shoemaker BA, Wang J, Wolynes PG. 1997. *Proc. Natl. Acad. Sci. USA* 94:777–82
185. Bohr HG, Wang J, Wolynes PG. 1994. In *Protein Structure by Distance Analysis*, ed. H Bohr, S Brunak, pp. 98–109. Amsterdam: IOS
186. Abkevich VI, Gutin AM, Shakhnovich EI. 1995. *Protein Sci.* 4:1167–77
187. Panchenko AR, Luthey-Schulten Z, Wolynes PG. 1996. *Proc. Natl. Acad. Sci. USA* 93:2008–13
188. Radford SE, Dobson CM, Evans PA. 1992. *Nature* 358:302–7
189. Radford SE, Dobson CM. 1995. *Philos. Trans. R. Soc. London Ser. B* 348:17–25
190. Saven JG, Wang J, Wolynes PG. 1994. *J. Chem. Phys.* 101:11037–43
191. Wang J, Saven JG, Wolynes PG. 1996. *J. Chem. Phys.* 105:11276–84
192. Zwanzig R. 1995. *J. Chem. Phys.* 103:9397–400
193. Wang J, Onuchic J, Wolynes PG. 1996. *Phys. Rev. Lett.* 76:4861–64
194. Gutin A, Abkevich V, Shakhnovich EI. 1996. *Phys. Rev. Lett.* 77:5433–36
195. Wang J, Plotkin SS, Wolynes PG. 1997. *J. Physique I* 7:395–421
196. Takada S, Wolynes PG. 1996. Los Alamos Cond-mat preprint 4701165
197. Schweizer K. 1973. *Physica Scripta T* 49:99–106
198. Roan J, Shakhnovich E. 1996. *Phys. Rev. E* 5:5340–57
199. Thirumalai D, Ashwin V, Bhattacharjee JK. 1996. *Phys. Rev. Lett.* 77:5385–88
200. Timoshenko EG, Kuznetsov YA, Dawson KA. 1996. *Phys. Rev. E* 54:4071
201. Takada S, Portman JJ, Wolynes PG. 1997. *Proc. Natl. Acad. Sci. USA* 94:2318–21
202. Jackson SE, Fersht AR. 1991. *Biochemistry* 30:10428–35
203. Fersht AR. 1995. *Proc. Natl. Acad. Sci. USA* 92:10869–73
204. Hopfield JJ. 1982. *Proc. Natl. Acad. Sci. USA* 79:2554–58
205. Ritter H, Martinetz T, Schulten K. 1992. *Textbook: Neural Computation and Self-Organizing Maps: An Introduction.* New York: Addison-Wesley. Revi. English ed.
206. Tanaka S, Scheraga HA. 1976. *Macromolecules* 9:945–50
207. Sippl MJ. 1990. *J. Mol. Biol.* 213:859–83
208. Casari G, Sippl MJ. 1992. *J. Mol. Biol.* 224:725–32
209. Bowie JU, Lüthy R, Eisenberg D. 1991. *Science* 253:164–70
210. Lüthy R, Bowie JU, Eisenberg D. 1992. *Nature* 356:83–85

211. Mirny LA, Shakhnovich EI. 1996. *J. Mol. Biol.* 264:1164–79
212. Miyazawa S, Jernigan RL. 1996. *J. Mol. Biol.* 256:623–44
213. Jernigan RL, Bahar I. 1996. *Curr. Opin. Struct. Biol.* 6:195–209
214. Goldstein RA, Luthey-Schulten Z, Wolynes PG. 1994. In *Proc. 27th Hawaii Int. Conf. on System Sciences.* pp. 306–15. Los Alamitos, California: IEEE Comput. Soci.
215. Maiorov VN, Crippen GM. 1992. *J. Mol. Biol.* 227:876–88
216. Kocher J, Rooman M, Wodak SJ. 1994. *J. Mol. Biol.* 235:1598–613
217. Rooman MJ, Wodak SJ. 1996. *Protein Eng.* 8:849–58
218. Hinds DA, Levitt M. 1996. *J. Mol. Biol.* 258:201–9
219. Kolinski A, Skolnick J. 1994. *Proteins* 18:338–52
220. Kolinski A, Skolnick J. 1994. *Proteins* 18:353–66
221. Jones D, Thornton J. 1993. *J. Comput. Aided Mol. Design* 7:439–56
222. Govindarajan S, Goldstein RA. 1996. *Proc. Natl. Acad. Sci. USA* 93:3341–45
223. Li H, Helling R, Tang C, Wingreen N. 1996. *Science* 273:666–69
224. Li H, Helling R, Tang C, Wingreen N. 1996. *J. Mol. Biol.* submitted
225. Chothia C. 1994. *Development* S:27–33
226. Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. *J. Mol. Biol.* 247:536–40
227. Jones D, Taylor WR, Thornton J. 1992. *Nature* 358:86–89
228. Goldstein RA, Luthey-Schulten ZA, Wolynes PG. 1996. In *New Developments in Theoretical Studies of Proteins,* ed. R Elber, pp. 359–88. Singapore: World Sci.
229. Koretke KK, Luthey-Schulten Z, Wolynes PG. 1996. *Protein Sci.* 5:1043–59
230. Wolynes PG. 1995. In *Proc. Symp. on Distance-Based Approaches to Protein Structure Determination II,* ed. H Bohr, S Brunak, J Keiding, pp. 3–15. New York: CRC
231. Haney P, Konisky J, Koretke KK, Luthey-Schulten Z, Wolynes PG. 1997. *Proteins: Struct. Func. Genet.* 28:117–30
232. Abkevich VI, Gutin AM, Shakhnovich EI. 1996. *Folding Design* 1:221–30
233. Gutin AM, Abkevich VI, Shakhnovich EI. 1995. *Proc. Natl. Acad. Sci. USA* 92:1282–86
234. Shakhnovich EI. 1994. *Phys. Rev. Lett.* 72:3907–10
235. Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA. 1995. *Proc. Natl. Acad. Sci. USA* 92:325–29
236. Deutsch JM, Kurosky T. 1996. *Phys. Rev. Lett.* 76:323–26