VOLUME 31 NUMBER 8

ACCOUNTS of CHEMICAL RESEARCH®

AUGUST 1998

Theory of Two-State Cooperative Folding of Proteins

MING-HONG HAO AND
HAROLD A. SCHERAGA*

Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853-1301

Received December 17, 1997

Introduction

Two of the fundamental problems in the theory of protein folding are how a protein folds to its native structure in a kinetically feasible time and whether the native state of the protein is stable thermodynamically under the folding conditions. These two problems are also the key issues encountered in designing a reliable model for folding proteins by computational means. Recent theoretical studies have achieved a better understanding of these problems. Remarkably, both the thermodynamic and the kinetic problems of protein folding can be resolved by following a consistent line of physical reasoning. Over the years, this line of theory has been expressed in the form of the principle of minimum frustration,[1,2] the topography of the folding funnel,[3−5] the pronounced energy minimum for the native structure,[6,7] and in the perspective of the statistical energy landscape of proteins.[8,9] In this article, we expose another aspect of the theory: namely, foldable protein models have the characteristics of two-state systems, and cooperativity is an essential condition of foldability of protein models. We will summarize the essential behavior of various protein-like models studied previously, describe the differences as well as the common physical basis of cooperative two-state folding in these theoretical models, and explain how force fields can be derived, based on the above knowledge, for protein models that fold quickly to unique native structures with the properties of realistic proteins. The present treatment correlates and extends previous expositions.[5,7,9] As the theory develops, it will become more concrete and computable, and, hopefully, can make more precise predications about practical problems.

To start with, we should clarify the physical context of the two-state folding of protein models. Experiments have shown that most single-domain globular proteins exhibit the character of a two-state or all-or-none folding/unfolding transition;[10] in multiple-domain proteins, the folding of individual domains can also be treated by the two-state model.[11] Any theory of protein folding must account for this essential feature.[7,12] In computer simulations, while noncooperative protein models can be folded to their native structures, such models usually involve short chains or ones that are dominated by local interactions. The most general protein models, which have the typical size and flexibility of real proteins and which fold fast to a stable native structure under folding conditions, appear to be two-state systems. While the apparent folding processes of a chain molecule can be deceptive, for example, a two-state system may fold in a downhill manner under certain conditions, the two-state protein model can be identified nonambiguously based on its intrinsic properties. In our treatment, it is based on the density of states of a protein system.

The Microcanonical Entropy Function

The density of states, $\Omega(E)$, is the number density of conformations as a function of the total energy, E, which

Ming-Hong Hao was born in Liaoning, China, in 1956, and received the B.A. and M.A. degrees from Liaoning and Nanjing Universities, respectively, in China. He completed his Ph.D. research at Rutgers University in 1989, taught as a lecturer at Shanxi University in 1990, and did postdoctoral work at the University of Alabama in Birmingham in 1991. Since 1992, he has been working in Harold Scheraga's lab at Cornell University and, currently, is a research associate. His main research interests are in the areas of computational methods for simulating biopolymers and statistical mechanics of protein folding.

Harold A. Scheraga was born in Brooklyn, NY, in 1921. He attended the City College of New York, where he received the B.S. degree, and went on to graduate work at Duke University, receiving the Ph.D. degree in 1946 and an Sc.D. degree (honorary) in 1961. He is now Todd Professor of Chemistry Emeritus at Cornell University. His research interests are in the physical chemistry of proteins and other macromolecules, the chemistry of blood clotting, and the structure of water and dilute aqueous solutions.

S0001-4842(96)00288-9 CCC: $15.00
Published on Web 05/15/1998

is treated as the energy of a protein conformation averaged over all solvation interactions. The density of states is more conveniently represented by its logarithm, $S(E) = \ln \Omega(E)$, termed the microcanonical entropy function. Given $S(E)$, one can calculate the relative free energy as a function of energy, $F(E) = E - TS(E)$, which defines the statistical probability of the state with energy E. Other canonical properties of the system can be calculated readily from these functions. The essential characteristic of a two-state system is that its entropy function $S(E)$ contains a concave segment.[13−15] Analytically, the concave segment of the entropy function is defined as a region of the function where the second derivative $d^2S(E)/dE^2$ is positive (with the generally valid condition that $S(E)$ is a continuously increasing function of E in physically relevant regions). When $S(E)$ contains a concave segment which lies between the native and non-native states, the free energy $F(E)$ will have two minima at the native and non-native states, respectively, at the folding temperature. This is the intrinsic condition for a two-state system. In infinite systems where two different phases coexist at the transition point, a two-state first-order transition is signaled by a linear segment of the microcanonical entropy curve.[16] However, for a finite protein, the two-state transition is defined by the free-energy barrier, which is associated with a concave segment of the entropy function.

The folding behavior of a protein is defined by its overall energy landscape. However, the energy landscape is a high-dimensional function which is difficult to access and utilize directly. One way to overcome this problem is to project the energy landscape into one- or low-dimensional functions. The microcanonical entropy function is a very useful one-dimensional mapping of the energy landscape. When a multidimensional energy surface is mapped into $S(E)$ by an integration, i.e.,

$$S(E) = \ln \int \delta(E - E') \, d\sigma_{E'}$$

where $\sigma_{E'}$ is the local surface element and the integral is evaluated over the complete energy surface, the local ruggedness of the energy landscape is usually averaged out. This leads to a loss of information about the detailed dynamics of protein folding. A complete treatment of the folding kinetics of proteins requires an analysis of the local ruggedness of the potential energy surface.[17] However, for good protein models that satisfy the principle of minimum frustration,[1,2] their folding transition can occur well above the glass transition temperature,[7−9] and the effects of a conformational diffusive process on folding may be treated as a prefactor in the kinetic rate expression.[9] For such good protein models, the one-dimensional microcanonical entropy function provides the essential information about the thermodynamics (the free energy function) and the kinetics (the folding barrier) of the folding transitions. With the advantage that it can be determined reliably by computational procedures (see below), the microcanonical entropy function provides a powerful tool for studying the problem of protein folding.

Two efficient procedures for determining the density of states of protein models are the Monte Carlo histogram (MCH) and the entropy sampling Monte Carlo (ESMC) methods. MCH[18] is based on the conventional MC algorithm. To sample all energy states, a number of conventional MC runs have to be carried out at different temperatures. In an MCH approach, the numbers of sampled conformations at different energy levels are saved as a histogram, $H(E)$; different histograms are then combined together in such a way as to minimize the statistical errors in individual histograms and to produce an optimal relative entropy function for the relevant energy region of a protein model. In the ESMC procedure,[19,20] the statistical weight for sampling an energy state is determined by $P(E) \propto \exp[-S(E)]$. Such a procedure enhances the sampling of the low-density (low-energy) states that would otherwise not be sampled sufficiently. The trial entropy function $S(E)$, which starts with an estimate of the true entropy function, is updated according to the energy histograms from a previous ESMC simulation, $S(E)_{new} = S(E)_{old} + \ln H(E)_{old}$. In this way, after a number of iterations, $S(E)$ approaches the correct microcanonical entropy. With the correct entropy function, the sampling probabilities of all relevant energy states in an ESMC simulation are equal; theoretically, one simulation can access all energy states below a cutoff energy level. The relative efficiency of the above two methods for determining the microcanonical entropy function depends on the nature of the protein model as well as on the conformational updating algorithm.[21,22]

Because proteins are complicated macromolecular systems, at present only simplified protein models can be analyzed in detail by the statistical-mechanical methods described above. Our insights about the cooperative nature of protein folding have been derived mainly from two types of lattice-chain protein models. The first model, termed Type I, is a cubic-lattice chain with contact interactions between noncovalent nearest-neighbor residues in the lattice. Because of its simplicity, this type of model has been studied extensively by both computer simulation and analytical treatments.[1,3−7,17,23−26] While such a "minimalist" model can capture a number of basic characteristics of proteins, it lacks the backbone interactions in polypeptides that lead to a preference for certain local conformations and hydrogen bonding which limits the folding pattern of proteins.[27] This concern is relieved by another protein model, termed Type II, which is a fine-grained lattice-chain model with specified local conformational preferences, backbone hydrogen bonding, and long-range side-chain contact interactions.[28−31] These two types of models represent different possible kinds of interactions or force fields for protein chains, and their microcanonical entropy functions have been determined to comparable accuracy.[14,15,21,22,32,33] It is significant that these two types of protein models reveal two limiting mechanisms of cooperative folding; therefore, it is hopeful that they cover, or can be used to conjecture, the folding behavior of other more realistic but more complicated protein models[34,35] whose statistical-mechanical proper-

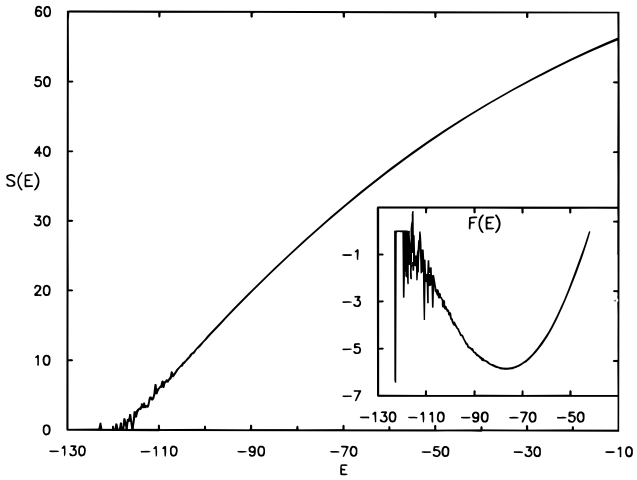

FIGURE 1. Relative microcanonical entropy of a Type I protein model. The entropy of the native state (the small peak at the lowest energy end) is arbitrarily set to unity. Inset: the relative free energy of the model as a function of the total energy at the folding temperature.

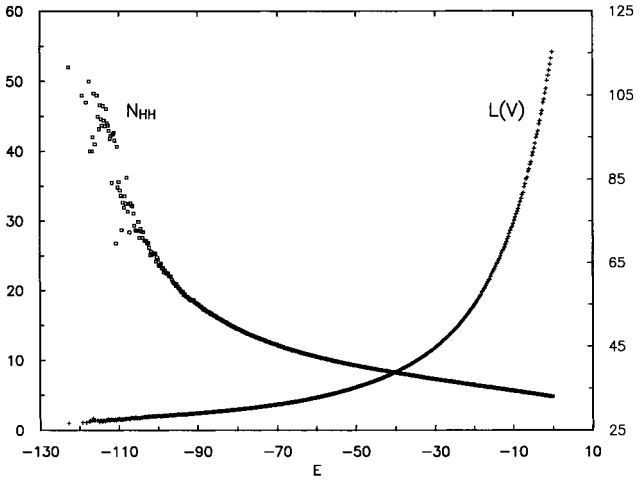

FIGURE 2. The average number of native contacts (N_{HH}, left axis) and the average volume [$L(V)$, right axis] of model I as a function of the total energy. $L(V)$ is calculated approximately as the volume of a sphere with a radius equal to the radius of gyration of the average conformation at the corresponding energy level.

ties could not be determined as completely and/or accurately as for the lattice models.

Thermodynamics of Two-State Protein Models

While the general character of two-state systems is well-known (see above), it is important to understand how such features are realized in concrete protein models and what their physical basis is. The above two types of models provide such information. Let us start with the Type I model; Figure 1 shows the microcanonical entropy curve for this model. The characteristics of this system are that there is a unique native state, whose energy is lowest compared to all other non-native states, and a concave segment in the entropy curve at the low-energy region [when $S(E)$ at that region is approximated as a smooth curve].[22,33] The two-state nature of this system is seen more clearly from the free-energy plot in the inset of the figure where there are two free-energy minima at the folding temperature. The characteristics of the entropy and free-energy curves shown in Figure 1 are representative of Type I protein models with good sequences and/or optimized interaction parameters.[5−7,17,24−26] The energy gap[5,6] between the native and non-native states in this model is correlated with the concave segment at the low end of the entropy curve; specifically, if there is an energy gap separating the native state from other high-energy non-native states, the lower end of the entropy function will be concave when the curve is smoothed.

The structural basis of the two-state behavior of Type I models is revealed from the variations of conformational properties, such as the average number of native contacts and the average volume of the chain conformation, as a function of the total energy.[33,36] The relevant information is shown in Figure 2. It is found that there are two main phases of conformational changes: In the high-energy region, the number of native contact is small while the expanded chain transforms quickly into a compact state when its energy decreases; in the low-energy region, in comparison, there is a rapid increase in the number of native contacts while the average volume of the chain remains small. The free-energy minimum in the *non-native* state of the Type I model, i.e., the high-energy region of Figure 2, can be attributed to a random collapse of the chain conformation which increases the chance of interactions among all residues and leads to a large decrease in energy. At a proper temperature, the decrease of energy compensates for the loss of configurational entropy due to volume reduction, so that the free energy reaches a minimum. The other free-energy minimum in the *native state* arises from the burst increase in the number of native contacts as seen in the low-energy region of Figure 2. The native contacts decrease the total energy much more than the random contacts in the non-native compact state; such a decrease in energy compensates for the unfavorable entropy of the ordered native state in comparison to the random mixing state, so that the free energy has another minimum. These are the structural bases of the two-state behavior of Type I models.

Because the interactions in the random collapsed state are basically similar among all sequences of sufficient length and general composition, all such sequences are expected to have a collapsed state in which the free energy is a minimum below a certain temperature. However, due to the bonding constraints of the chain in a finite protein, only *good* sequences can form a large number of native contacts simultaneously to reduce the energy sufficiently, making the native structure a free-energy minimum and giving rise to the two-state character. Therefore, the behavior shown in Figures 1 and 2 is the result of optimized sequences/interactions in Type I models.

We now turn to Type II models with two-state characteristics. Figure 3 shows the microcanonical entropy curve of such a protein model.[15] The essential characteristic of this system is that there is a broad concave segment in the intermediate energy region of the micro-

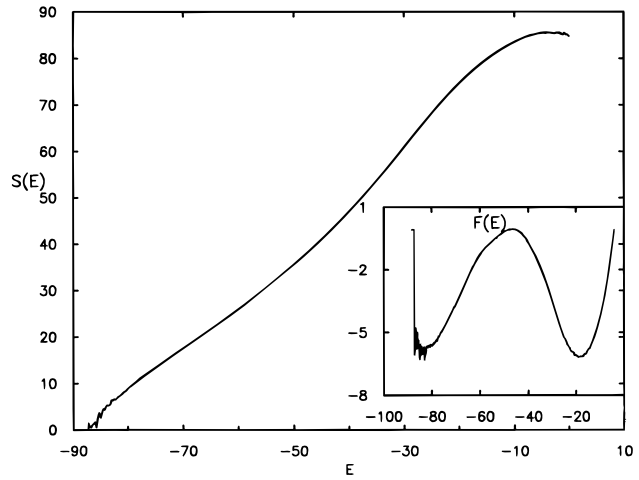

FIGURE 3. Relative microcanonical entropy of a Type II protein model. The entropy of the native state (the small peak at the lowest-energy end) is arbitrarily set to unity. Inset: the relative free energy of the model as a function of the total energy at the folding temperature.

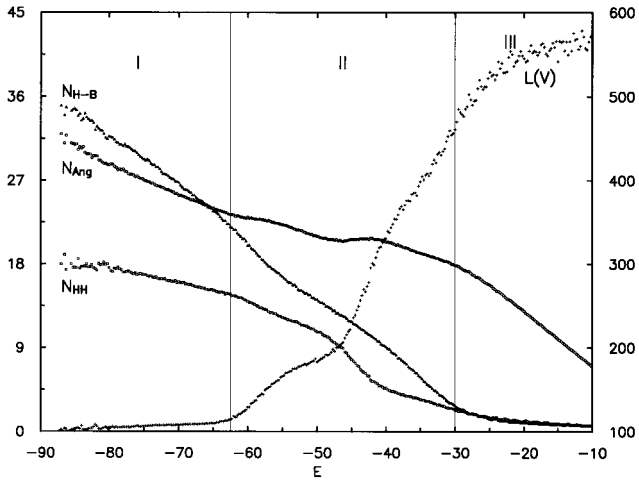

FIGURE 4. The average number of side-chain native contacts (N_{HH}), of native virtual-bond angles (N_{Ang}), and of hydrogen bonds (N_{H-B}), left axis, and the average volume [$L(V)$], right axis, at different energy levels. $L(V)$ is defined as in Figure 2 but in units of 10^{-1} (lattice unit)^3.

canonical entropy curve, in addition to a unique structure with lowest energy. The corresponding free energy function at the folding temperature is shown in the inset of the figure, where there are two comparable free-energy minima in the non-native (high-energy) and the native (low-energy) regions, respectively. In this system, the native state as defined by the free-energy minimum is not a single structure; it consists of an ensemble of conformations with similar overall structure but varying energies. As a result, the free-energy minima of the native and non-native states are relatively symmetric. The behavior shown in Figure 3 is representative of this type of model with various optimized sequences,[14,15,32] as well as with more sophisticated interaction potentials.[31]

To describe the molecular origin of the two-state character of the Type II models, Figure 4 shows the variations of the conformational properties with the total energy, including the average volume, the average number of native backbone virtual-bond angles, hydrogen bonds, and native contacts that are all involved in the potential function of this model.[36] The free-energy minimum in the *non-native* state of this model is associated with the formation of locally structured units, as indicated by the rapid increase in the number of native bond angles (N_{Ang}) in the high-energy region of Figure 4. The reason for this is that the energy of the system is greatly reduced when many local chain conformations adopt their preferred states while the conformational entropy of the chain remains high because the average volume of the chain is large and the global conformation is still flexible; hence, the free energy can reach a minimum. In comparison, the free energy minimum in the *native state* results from the large energy decrease due to the formation of long-range hydrogen bonding (N_{H-B}) and side-chain native interactions (N_{HH}), as indicated by the variations of the above properties in the low-energy region of Figure 4. At the folding temperature, these energy decreases compensate for the loss of entropy in the native state in comparison to the non-native state, so that the free energy reaches another minimum.

It has been found[15,32] that the two-state characteristic of Type II models is the result of consistency among short-range energies (local conformational preferences) and long-range interactions (hydrogen bonding and side-chain interactions) in the native structures. When there are conflicts among the different energy components, as when the sequences are chosen randomly, the system loses its two-state behavior.[15] Increases in the strength of local energies make the concave segment of the entropy curve more stretched out and the energy minimum of the native state more pronounced.[32]

The two-state characteristics of Type I and II models can be described by a simple mean-field theory,[37] which helps to highlight the differences between these two types of models quantitatively. In this theory, the total energy of a protein in a given state is expressed as a series expansion in the number, m, of residues that are in their local native states,

$$E_{\text{total}}(m) = \epsilon_0 + \epsilon_1 m + \epsilon_2 m^2/N + \ldots$$

where ϵ_0 is a constant reference energy, ϵ_1 and ϵ_2 are, respectively, the mean-field single-residue energy and residue−residue pairwise interaction energy, and N is the total number of residues. In this mean-field treatment, the entropy of a state at a given m is calculated as

$$S(m) = \nu^{(N-m)} N!/m!(N-m)!$$

where ν is the number of non-native states for each residue in the system. With a proper choice of the energetic and conformational parameters, the above simple mean-field formulation can produce the microcanonical entropy functions of the *various* protein models; the calculation of the thermodynamic and kinetic properties of the system can then easily follow.[37,38] The most interesting result of this formulation is that the Type I model is characterized by a mean-field energy expression

with only the single-residue energy term, plus an energy gap between the native state and other states, while the character of Type II models has to be described by an energy expression with *both* the single-residue and double-residue energy terms. This result indicates the differences in the dominant interactions in these two types of models. The above analysis also showed that the Type I model has the characteristics of simple random-energy systems with a single Gaussian-like distribution of energies;[7,39] the stability of the native state and cooperativity of folding of this model arise primarily from the energy gap between the native and the non-native states.[38] In the Type II model, on the other hand, the attributes of random-energy systems superimpose on the tendency of minimum frustrations in the conformational states,[2,32] as indicated by the correlations among the residues that adopt their native states. In a more detailed mean-field analysis in which each state *m* is provided with a Gaussian-like energy distribution,[2] it was found that there is an automatical jump in the fraction of native residues in the transition state for two-state Type II models.[32] The physical implication of such a jump in the order parameter is that the dominant mean-field interactions of the system switch from single-residue interactions to residue−residue pairwise interactions.[37] This behavior is quite different from that of Type I models.

Free-Energy Barriers and Folding Kinetics

For good protein models whose folding funnels are relatively smooth, the elementary rate of diffusion or transmission of chain conformation over the energy landscape may be treated approximately as a constant above or near the folding temperature.[37,38] The overall folding kinetics of a two-state system is then controlled by the free-energy barrier. Here, we examine the origins of the free-energy barriers in the two protein models.

In Type I models (see Figures 1 and 2), the conformational transition to the native state occurs in a compact state, which has a favorable mixing entropy due to random contacts among the residues. The initial conversion of a random state to an ordered state dramatically reduces the mixing entropy, causing an increase in the free energy and resulting in a free-energy barrier as seen in the inset of Figure 1. In other words, the free-energy barrier arises in forming the native-contact nucleus. However, once a native nucleus is formed, many residues can condense simultaneously onto the native nucleus to form native contacts. This quickly reduces the energy and compensates for the loss of entropy, leading the protein to the free-energy minimum in the native state. The term "cooperativity" refers to the behavior of the conformational change after the protein has passed over the free energy barrier. In this sense, we can attribute the origin of cooperativity of Type I models to "a concerted action of a large number of residues forming native contacts simultaneously in a compact conformation".[36]

The folding of Type I protein models can be summarized in a three-stage process.[6,34] First, the chain collapses rapidly into a compact conformation, it then follows a slow searching process in the compact state, and, finally, the chain converts to the native state in a cooperative manner. In some studies,[4,9] the compact state of the Type I model has been ascribed to a molten globule state. However, since the conformational distribution in this compact state is rather random, such a state is perhaps quite different from the experimentally observed molten globule state in which there is a significant number of native contacts and much of the native secondary structure.

In Type II models (see Figures 3 and 4), the conformational transition as well as the free-energy barrier occur in a relatively open state. The barrier arises as a result of initial formation of native contacts and/or interactions among locally structured units, which greatly constrains the global freedom of the chain and reduces the conformational entropy, giving rise to a free-energy barrier. To reduce the entropy cost, it is more favorable for the initial formation of native contacts to occur between structured units that are close to each other along the chain. Once the initial correct contacts are made, however, further condensation of other locally structured units onto the native nucleus involves less of an entropy loss, and a large decrease of energy, because the latter added units can form native interactions with several other structured units that are already in the native cluster. These combined effects lead to a decrease in the free energy, guiding the chain to assemble quickly into the native structure at the free-energy minimum.

The folding transition of Type II models can also be described as a three-stage process.[36] First, local chain conformations form uncorrelated segments of locally structured units; these are not permanent and can form or disrupt reversibly. Second, two or a small number of locally structured units form a partial native cluster or nucleus; this is a high free-energy state. Finally, other locally structured units condense onto the native core in a highly cooperative manner. The movements of locally structured units in forming the native structure resemble an "on site" assembly process[28] but with a certain degree of random condensation.[36] This type of collective orientational arrangements of locally structured units can occur in proteins with all types of local secondary structures. The high cooperativity in the folding of the Type II model, which is the basis of a two-state transition, results from the simultaneous condensation of many structured units onto the native nucleus.

The Type I and II models represent two limiting behaviors of cooperative folding in protein-like polymers. Type I models reflect the characteristics of "soft" polymers which lack backbone interactions and local conformational preferences. Because the only forces that drive the folding of this type of model are the contact interactions among residues, a cooperative transition of such a chain model can occur only in a compact state where all the residues have maximum opportunity for contact interactions. In comparison, Type II models reflect the characteristics of "stiffer" polymers with *specific* backbone

interactions. The folding of this type of model is driven by heterogeneous forces: the local interactions not only lower the conformational energy but also trigger changes in global conformation; the long-range interactions between structured units are cooperative and orientationally selective. The differences between the driving forces for folding the two types of protein models lead to different folding behavior. In Type I models, the folding transition starts with a random collapse, and follows a large number of paths in conversion to the native state. In Type II models, the folding process initiates by forming locally structured units, and follows a smaller number of folding pathways. The folding of real proteins shows the behavior of both types of models: it involves the features of "on site" assembly as in Type II models and elements of random condensation as in Type I models. For example, the folding transition of an optimized 27-mer cubic-lattice chain has been mapped to the folding of a 60-residue helical protein,[4] indicating the random condensation feature of the latter system. On the other hand, experimentally observed formation of locally structured units, such as α-helical segments or β-strands in the earlier stages of folding of real proteins,[40] resembles the folding behavior of Type II models.

It is clear that each of the Type I and II models presents a simplified theoretical version of the two-state folding of real proteins. An important question is how much these two models reveal the whole story of two-state dynamics and cooperativity. In terms of the topography of two-state protein systems as characterized by a free energy barrier between the native and non-native states, the two types of models described here represent two limiting cases; that is, the concave segment of the microcanonical entropy function may appear either at the low-energy end, as in the Type I model, or in the intermediate energy region, as in the Type II model. In terms of the interaction schemes or force fields that lead to two-state characteristics, however, the protein models that have been studied so far perhaps are not exhaustive. Some interactions, such as multibody interactions, could lead to enhanced cooperativity. In one study with a cubic-lattice chain model,[33] the inclusion of solvation energy for individual residues, which is a kind of multibody interaction, still produces a microcanonical entropy curve such as a Type I model. But in another study with a fine-grained lattice model, the inclusion of specific multibody interactions enhances the characteristics of the Type II model in that system.[31] In general, it appears that the types of interactions that help the initial formation of local secondary structures promote the behavior of a Type II model. But the quantitative contributions of chain rigidity, specific multibody interactions, and the addition of side chains to the degree of cooperativity in protein folding require further investigations. Another relevant question is which of the Type I and II models is the preferred one for theoretical studies of protein folding. The Type II model is obviously more realistic than the Type I model because the former includes the critical feature of a protein backbone; therefore, a Type II model would have to be used in modeling real proteins. On the other hand, Type I models are much easier for computational studies and simpler for statistical-mechanical analysis; hence, they can be treated more rigorously. Therefore, the Type I model provides a solid starting point upon which we can build more sophisticated and realistic protein models.

The two-state mechanism solves the thermodynamic as well as the kinetic problems of protein folding: First, at the folding temperature, the native state is one of the free-energy minima, so that the native structure is stable. Second, kinetic traps are minor at the folding temperature because the statistical probability of low-energy non-native states is much smaller than that of the native state, so that folding to the native state is dominant. The folding is relatively fast at the two-state folding temperature because the thermal fluctuations of the system can easily overcome local ruggedness of the potential energy surface so that the elementary diffusion rate of the chain conformation is high. The cost for two-state folding is the hurdle of a free-energy barrier, but the benefits of this mechanism are greater than that of other mechanisms such as continuous folding or search at lower temperature where kinetic traps are significant. In simulation studies of both Type I and II models,[14,15,21,22] it has been found that, for reasonably long chains, a strong two-state characteristic is required for sufficiently fast folding. The general trend is that, the larger the system, the stronger are the two-state characteristics, such as a larger energy gap and a longer concave segment in the microcanonical entropy curve, which are required for the model to fold to the native state in a reasonable time.

Implications for the Design of Good Protein Models

A major goal of theoretical studies of protein folding is to develop proper computational models for theoretical folding of realistic proteins. Any good protein model must resolve both the thermodynamic and kinetic problems of folding. Knowledge of the common characteristics of two-state systems has led to a general strategy for optimizing the force field for protein models. To achieve the essential condition of a two-state system, i.e., a large energy gap between properly defined non-native and native states,[6,7,39] a straightforward method is to adjust the energy parameters of a properly chosen potential function so as to increase the energy differences between the non-native states and native states of given protein models. A more effective object function for optimization is the ratio of the difference between the average energy of the non-native states and the energy of the native state to the standard deviations of the energies of the non-native conformations.[41]

Recently, a number of algorithms have been proposed for optimizing the energy parameters according to the above idea. Goldstein et al.[41] developed an analytical formula for maximizing the above normalized energy difference. Mirny and Shakhnovich[42] proposed an alternative averaging scheme for evaluating the sum of object

functions in optimization of the energy parameters. To obtain sufficiently good energy parameters for folding large protein models, the definition of the non-native states is critical because some non-native states are correlated with the variations of the energy parameters. To overcome this problem, we have developed an optimization procedure that uses Monte Carlo sampling to generate the non-native ensemble for each set of energy parameters and employs an iterative process to obtain a set of self-consistent optimized energy parameters.[21,22,33] An alternative iterative scheme for optimizing energy parameters has also been proposed by Deutsch and Kurosky.[43] Finally, Thomas and Dill[44] developed a procedure that optimizes the energy parameters by maximizing the Boltzmann probabilities of the native structures of training proteins over the sum of Boltzmann probabilities of the ensemble of all conformations. The basis of this algorithm is to minimize the energies of native structures, which is clearly in line with the above general idea. However, the above procedures have been mostly successful in systems with the characteristics of the Type I model discussed above. Further work is required to develop an equally successful procedure for optimizing the force fields of the Type II models.

Summary

This article summarizes our theoretical understanding of the basic and simplest form of protein folding, i.e., the two-state transition. It is clear now why two-state conditions resolve both the thermodynamic and kinetic problems for protein folding. Two limiting forms of cooperative folding in protein models are described; they define a physical range in which we may expect the molecular mechanism of the folding transitions of real proteins to lie. While many of the observations were made here on the basis of highly simplified protein models, the characteristics of the microcanonical entropy functions of the two types of protein models are very general, and provide two limiting references for analyzing the behavior of general protein models. The most exciting aspect of the above knowledge is that it can guide the development of good and realistic protein models that fold fast to correct native structures.

This work was supported by grants from the National Science Foundation (MCB95-13167) and the National Institutes of Health (GM-14312).

References

(1) Gō, N. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183−210.
(2) Bryngelson, J. D.; Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 7524−7528.
(3) Leopold, P. E.; Montal, M.; Onuchic, J. N. Protein folding funnels: A kinetic approach to the sequence-structure relationship. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 8721−8725.
(4) Onuchic, J. N.; Wolynes, P. G.; Luthey-Schulten, Z.; Socci, N. D. Toward an outline of the topography of a realistic protein-folding funnel. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 3626−3630.
(5) Dill, K. A.; Chan, H. S. From Levinthal to pathways to funnels. *Nat. Struct. Biol.* **1997**, *4*, 10−19.
(6) Sali, A.; Shakhnovich, E. I.; Karplus, M. Kinetics of protein folding. A lattice model study of the requirements for folding to the native state. *J. Mol. Biol.* **1994**, *235*, 1614−1636.
(7) Shakhnovich, E. I. Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **1997**, *7*, 29−40.
(8) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **1995**, *21*, 167−195.
(9) Onuchic, J. N.; Luthey-Schulten, Z.; Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **1997**, *48*, 545−600.
(10) Privalov, P. L. Stability of proteins. Small globular proteins. *Adv. Protein Chem.* **1979**, *33*, 167−241.
(11) Privalov, P. L. Stability of proteins which do not present a single cooperative system. *Adv. Protein Chem.* **1982**, *35*, 1−104.
(12) Poland, D. C.; Scheraga, H. A. Statistical mechanics of noncovalent bonds in polyamino acids. IX. The two-state theory of protein denaturation. *Biopolymers* **1965**, *3*, 401−419.
(13) Gō, N. Theory of reversible denaturation of globular proteins. *Int. J. Pept. Protein Res.* **1975**, *7*, 313−323.
(14) Hao, M.-H.; Scheraga, H. A. Monte Carlo simulation of a first-order transition for protein folding. *J. Phys. Chem.* **1994**, *98*, 4940−4948.
(15) Hao, M.-H.; Scheraga, H. A. Statistical thermodynamics of protein folding: sequence dependence. *J. Phys. Chem.* **1994**, *98*, 9882−9893.
(16) Huller, A. First-order phase transitions in the canonical and the microcanonical ensemble. *Z. Phys. B* **1994**, *93*, 401−405.
(17) Socci, N. D.; Onuchic, J. N.; Wolynes, P. G. Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **1996**, *104*, 5860−5868.
(18) Ferrenberg, A. M.; Swendsen, R. H. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.* **1989**, *63*, 1195−1198.
(19) Berg, B. A.; Celik, T. New approach to spin-glass simulations. *Phys. Rev. Lett.* **1992**, *69*, 2292−2295.
(20) Lee, J. New Monte Carlo algorithm: entropy sampling. *Phys. Rev. Lett.* **1993**, *71*, 211−214; Erratum, **1993**, *71*, 2353.
(21) Hao, M.-H.; Scheraga, H. A. How optimization of potential functions affects protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 4984−4989.
(22) Hao, M.-H.; Scheraga, H. A. Optimizing potential functions for protein folding. *J. Phys. Chem.* **1996**, *100*, 14540−14548.
(23) Chan, H. S.; Dill, K. A. Polymer principles in protein structure and stability. *Annu. Rev. Biophys. Biophys. Chem.* **1991**, *20*, 447−490.
(24) Shakhnovich, E. I. Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett.* **1994**, *72*, 3907−3910.
(25) Dill, K. A.; Bromberg, S.; Yue, K.; Fiebig, K. M.; Yee, D. P.; Thomas, P. D.; Chan, H. S. Principles of protein folding - -A perspective from simple exact models. *Protein Sci.* **1995**, *4*, 561−602.

(26) Abkevich, V. I.; Gutin, A. M.; Shakhnovich, E. I. Impact of local and nonlocal interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.* **1995**, *252*, 460−471.
(27) Honig, B.; Cohen, F. E. Adding backbone to protein folding: why proteins are polypeptides. *Folding Design*, **1996**, *1*, R17−R20.
(28) Skolnick, J.; Kolinski, A. Simulations of the folding of a globular protein. *Science* **1990**, *250*, 1121−1125.
(29) Kolinski, A.; Godzik, A.; Skolnick, J. A general method for the prediction of the three-dimensional structure and folding pathway of globular proteins: application to designed helical proteins. *J. Chem. Phys.* **1993**, *98*, 7420−7433.
(30) Kolinski, A.; Galazka, W.; Skolnick, J. Computer design of idealized β- motifs. *J. Chem. Phys.* **1995**, *103*, 10286−10297.
(31) Kolinski, A.; Galazka, W.; Skolnick, J. On the origin of the cooperativity of protein folding - Implications from model simulations. *Proteins* **1996**, *26*, 271−287.
(32) Hao, M.-H.; Scheraga, H. A. Statistical thermodynamics of protein folding: comparison of a mean-field theory with Monte Carlo simulation. *J. Chem. Phys.* **1995**, *102*, 1334−1348.
(33) Hao, M.-H.; Scheraga, H. A. On foldable protein-like models; a statistical-mechanical study with Monte Carlo simulations. *Physica A* **1997**, *244*, 124−146.
(34) Camacho, C. J.; Thirumalai, D. Kinetics and thermodynamics of folding in model proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6369−6372.
(35) Boczko, E. M.; Brooks, C. L. First-principles calculation of the folding free-energy of a three-helix bundle protein. *Science* **1995**, *269*, 393−396.
(36) Hao, M.-H.; Scheraga, H. A. Molecular mechanisms for cooperative folding of proteins. *J. Mol. Biol.*, in press.
(37) Hao, M.-H.; Scheraga, H. A. Characterization of foldable protein models: thermodynamics, folding kinetics, and force field. *J. Chem. Phys.* **1997**, *107*, 8089−8102.
(38) Zwanzig, Z. Simple model of protein folding kinetics. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9801−9804.
(39) Shakhnovich, E. I.; Gutin, A. M. Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 7195−7199.
(40) Kim, P. S.; Baldwin, R. L. Intermediates in the folding reactions of small proteins. *Annu. Rev. Biochem.* **1990**, *59*, 631−660.
(41) Goldstein, R. A.; Luthey-Schulten, Z. A.; Wolynes, P. G. Optimal protein-folding codes from spin-glass theory. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 4918−4922.
(42) Mirny, L. A.; Shakhnovich, E. I. How to derive a protein folding potential? A new approach to an old problem. *J. Mol. Biol.* **1996**, *264*, 1164−1179.
(43) Deutsch, J. M.; Kurosky, T. New algorithm for protein design. *Phys. Rev. Lett.* **1996**, *76*, 323−326.
(44) Thomas, P. D.; Dill, K. A. An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 11628−11633.

AR960288Q