Annu. Rev. Biophys. Biomol. Struct. 2001. 30:361–96

PROTEIN FOLDING THEORY: From Lattice to All-Atom Models

Leonid Mirny and Eugene Shakhnovich
Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138; e-mail: shakhnovich@chemistry.harvard.edu; leonid@origami.harvard.edu

Key Words nucleation, folding nucleus conservation, molecular dynamics, structure prediction, folding nucleus prediction

■ **Abstract** This review focuses on recent advances in understanding protein folding kinetics in the context of nucleation theory. We present basic concepts such as nucleation, folding nucleus, and transition state ensemble and then discuss recent advances and challenges in theoretical understanding of several key aspects of protein folding kinetics. We cover recent topology-based approaches as well as evolutionary studies and molecular dynamics approaches to determine protein folding nucleus and analyze other aspects of folding kinetics. Finally, we briefly discuss successful all-atom Monte-Carlo simulations of protein folding and conclude with a brief outlook for the future.

CONTENTS

1056-8700/01/0610-0361$14.00

INTRODUCTION

Although no method exists that can reliably predict the native structure of a protein from its sequence, our understanding of mechanisms that govern protein folding has progressed considerably. Such progress was achieved by means of intensive experimental and theoretical studies of a broad class of small proteins and simple generic lattice and off-lattice models.

In theory, major insights came from folding simulations of simplified lattice and off-lattice models. Simplicity and computational efficiency of these models made it possible to simulate thousands of folding-unfolding events to obtain a detailed statistical description of the folding process in model proteins. Furthermore, it was also possible to model the evolution of proteins under various selective pressures. Although these models do not match the full complexity of real protein architecture, they capture a core aspect of the physical protein folding problem: Both real proteins and simplified lattice and off-lattice proteins find a conformation of the lowest energy out of an astronomically large number of possible conformations without prohibitively long exhaustive search. By simulating folding and evolution of simple model proteins, theoreticians gained insights into general properties of amino acid sequences that are required to provide stability and fast folding to model proteins. First, to exhibit cooperative folding transition and to fold fast, the native structure must be a pronounced energy minimum separated from the rest of the structure by a large energy gap. Second, selected protein sequences fold by nucleation mechanisms whereby a small number of residues (folding nucleus) need to form their native contacts in order for folding reaction to proceed fast into the native state.

As always, important developments create new challenges for theoretical and experimental research:

1. Can one predict/rationalize which residues contribute most to the folding nucleus given the sequence and the native structure of a protein?
2. How can one predict the stability and folding rate of a protein? At least, how can one predict changes in the stability and folding rate upon single mutation?
3. What is the evolution of the folding nucleus and folding kinetics in general? Are nucleation residues under stronger evolutionary pressure than the rest of the protein core?

A number of recent studies addressed these questions using a variety of techniques. In this review, we survey many of those recent works and emphasize their strong and weak points.

The structure of this review consists of the following: First, we introduce basic concepts essential for understanding further material. Next, we turn to discussion of recent theoretical studies and some experimental works. Finally, we summarize major conclusions and directions of further research that seem to us most important for understanding protein folding.

BASIC CONCEPTS

Cooperativity: Two-State Thermodynamics and Kinetics of Protein Folding

The concept of cooperativity is one of the most basic and fundamental for protein folding studies. Privalov and coworkers (96) applied the van't-Hoff criterion to evaluate the cooperativity of denaturational transitions for several proteins. They found, with a high degree of accuracy, that small single-domain proteins fold like two-state systems with only folded and unfolded states being (meta) stable, whereas all intermediate—partly folded—states are unstable (Figure 1). More recently, Jackson & Fersht (51) showed that folding of a small protein, chymotripsin inhibitor 2 (CI2), can be considered also as a kinetically two-state process in which partly folded states are not significantly populated in the process of folding. Subsequently, many more proteins were found to fold kinetically and thermodynamically as two-state systems. A list of such proteins as well as the data on their thermodynamics and kinetics are presented in an extensive review by Jackson (50).

The discovery of remarkable cooperativity of protein folding [akin to first-order phase transition for a finite system (109)] inspired many theoretical and experimental studies aimed at explaining it. Earlier models (partly reviewed in 52) considered factors such as polymer collapse (98), side-chain packing (109), directional or three-body interactions (47), or special folding pathways (23) as possible explanations.

More recently, cooperativity of protein folding received detailed explanation within analytical heteropolymer theory. Phenomenological (10) and microscopic statistical-mechanical theories (89, 100, 102, 110, 111) converged on the view that sequences that had undergone evolutionary selection fold cooperatively, whereas random sequences do not. Several simulation studies support this view (1, 45, 46, 103).

In a recent analysis (13, 54), Chan and coworkers analyzed cooperativity in various types of lattice models and found that of all studied sequence models, only three-dimensional models with twenty types of amino acids (Chan studied the same three-dimensional 36-mer sequence as was used in reference 42) feature folding cooperativity comparable to that of natural proteins. Two-dimensional models and hydrophobic polar (HP) models were found to be far less cooperative than real proteins. These results vindicate earlier detailed predictions from analytical models (3, 89, 100, 104).

The requirement of protein-like cooperativity narrows down the selection of viable models to study folding kinetics.

Transition State Ensemble and Folding Nucleus

Consider the folding of a protein that has two-state kinetics. In this case, folding proceeds through a single free-energy barrier. The height of the transition state, $\Delta G_{\ddagger-D}$, controls the rate of folding:

$$k_f = C \exp\left(-\frac{\Delta G_{\ddagger-\mathrm{D}}}{RT}\right),$$

where C is a constant. Approximate best fit experimental value of $C = 10^6\ \mathrm{s}^{-1}$ (36).

The transition state is not a single protein conformation. Rather, it is an ensemble of conformations (transition state ensemble, TSE) (104). Correspondingly, it is the free energy of the TSE, $\Delta G_{\ddagger-D}$, that determines folding rate (the number of conformations constituting the TSE is as important in determining protein folding rate as energy of the TSE).

The concept of TSE is a natural generalization of the concept of the transition state in chemical kinetics to protein folding with important consideration that folding occurs in a highly multidimensional space. Simple chemical reactions can be characterized by one (or very few) reaction coordinates (RC). The unique role of the reaction coordinate in chemical kinetics is that it provides a relation between the structure of the reagents (coordinates of the nuclei) and the time course of the reaction. Specifically, the value of RC can serve as a predictor of the transmission coefficient for the reaction: The top of the barrier separates the region where the flux is toward the products from the region where the flux is toward the reactants. In a simple chemical reaction described by a one-dimensional RC, the separation occurs at one particular value of the RC that is the transition state. In the case of complex protein folding reaction, no single RC is found at this point (see below; 30). Nevertheless, the concept of TSE, which is more general than RC, still applies. Indeed, the signature of a transition state in chemical kinetics is that the transmission coefficient for reactions that originate from TS is one half. If one imagines a statistical ensemble of reacting molecules, each starting a reaction from the transition state configuration, half of the molecules in the ensemble transform fast to products, and half of them transform back to reactants. This view can be generalized to multidimensional protein folding reactions: The TSE can be defined as a set of conformations such that folding trajectories starting from each of them have probability $p_{fold} = {}^1/_2$ to reach fast and downhill folded state before unfolding, and $p = {}^1/_2$ to reach the unfolded state before folding. Apparently the TSE is a separatrix in multidimensional conformational space that separates the basin of attraction for the folded state from that of the unfolded state. The folded basin of attraction (called postcritical in reference 2) consists of conformations that are committed to fast folding: Every folding trajectory starting from any conformation belonging to the postcritical ensemble reaches the folded state fast prior to unfolding via directed downhill motion in configurational space. It is very important to note that folding dynamics from postcritical conformations (PCCs) are qualitatively different from those starting from any conformation before the transition state. In the former case, a steady descent to the native state occurs

(Figure 1), whereas in the latter case, the dynamics feature a two-state character: Most of the time the chain appears to be making stochastic fluctuations with no apparent progress toward the folded state until at some moment it passes one of the conformations belonging to TSE and then rapidly jumps to the native state [in the time scale of total folding simulation, the descent from a TSE conformation to the native state looks like a jump, whereas on a finer time scale it represents a gradual biased descent (Figure 1)].

The nucleation concepts of protein folding kinetics were proposed and tested in the context of lattice models (2). The postcritical set of conformations for a simple 36-mer protein was determined, and it was directly verified that simulations that start from any of the conformations from this postcritical set indeed reach the native state via directed dynamics that no longer involve crossing of a major free-energy barrier (Figure 1). Furthermore, it was shown for the same model that dynamics starting from any PCC remains fast even at very low temperature (1), in contrast to dynamics that starts from an arbitrary non-PCC. This is the most direct verification that dynamics from postcritical conformations no longer involves major barrier crossing, which is energetic at low temperatures (1).

Which features distinguish conformations belonging to TSE and postcritical conformations from all other conformations? The nucleation theory suggests that fast-folding proteins have a small substructure (set of interactions) common to most of the conformations constituting the TSE. This substructure is called a folding nucleus (Figure 1). Stabilization of the folding nucleus lowers free energy of the TSE, $\Delta G_{\ddagger - D}$. This factor accelerates folding. Along the same line of arguments, destabilization of the folding nucleus leads to slower folding.

In most studied proteins and protein models, conformations in the TSE feature a certain set of native interactions between residues. Hence, a conformation with an assembled nucleus may look like a distorted native conformation, sometimes with large unstructured loops. The nucleus corresponds to a cluster of interacting residues that brings together different parts of the chain. Some of those residues are usually located far from each other along the sequence and form large unstructured loops when they are brought together to build the nucleus. Some proteins, however, have a nucleus with a partially formed secondary structure. In these cases, along with long-range interactions, short-range interactions that stabilize formed secondary structures contribute to the stability of the nucleus. The nucleation mechanism does not require secondary structure elements to be formed before the transition state is reached. In contrast to diffusion-collision and hierarchical folding models, secondary structure may be formed simultaneously with the folding of tertiary structure.

Several mechanisms alternative to nucleation were discussed in recent literature.

Folding Funnels

Originally, "folding funnels" were introduced (67) to indicate a special requirement of kinetic accessibility for a few viable native structures in 27-mer lattice models. However, subsequent lattice simulations did not support the view that only special

structures are kinetically accessible (simulations presented to support this view did not appear to be statistically significant). Correspondingly, the concept of the folding funnel has transformed into an intuitive reflection of the fact that the native state of a protein represents deep energy minimum; that is, most contacts in the native state are stabilizing.

This view makes it natural to explain the solution of the folding problem by the energetic bias that rewards every move towards the native state (the energetic funnel) (14). While such a landscape picture has some appeal due to its simplicity and artistic beauty, the funnel concept and related pictorial landscapes may be misleading. In fact, such a view contradicts experimentally observed exponential folding kinetics (7). The funnel picture also contradicts numerous lattice simulations where direct descent to the native state occurs at the very late (in time) stage of folding trajectory after the transition state is passed.

The major weakness of the funnel concept is that it downplays the importance of conformational entropy in the folding kinetics. It is the free energy, not the energy, that determines the direction and the timecourse of the protein folding reaction. The folding process is driven by opposing thermodynamic forces: the conformational entropy and the energy (enthalpy) of the residue-residue interactions. While the native state has the lowest energy, it also has the lowest conformational entropy. Before the transition state is reached, the conformational entropy dominates (i.e., formation of a new native interaction leads to a greater loss of the entropy than gain in the energy) and the process goes up-hill in free energy. After the transition state is passed, the process is dominated by the energy; that is, the energy gained upon formation of a new native interaction is greater than the loss of the conformational entropy. The process goes down-hill in free energy. Conformational entropy is the crucial component of the free energy balance in protein folding that metaphoric landscape pictures fail to take into account.

Diffusion-Collision Mechanism

A possible kinetic mechanism alternative to nucleation is suggested by framework or diffusion-collision models (DCM). These hierarchical models stipulate that folding starts from the formation of local stable structural elements, which serve as preformed units for subsequent stages of folding (6, 53).

The key difference between the nucleation mechanism and the hierarchical or DCM process is that they provide the qualitatively different predictions concerning the effect of mutations on folding rate. The DCM predicts that stabilization of any local secondary structures element (an α-helix or β-strand) will always lead to acceleration of folding. In contrast, the nucleation mechanism predicts that the strength of tertiary contacts formed in the transition state may be primary determinants of the folding rate (49, 14). According to the nucleation mechanism, stabilization of local structure accelerates folding rates only if a particular element is structured in the TSE.

The relative importance of secondary versus tertiary interactions in determining folding rates was studied experimentally for a number of proteins, including the

activation domain of carboxypeptyidase A (CPA) (84), CheY (84), the helical coiled-coil GCN4 (32, 82, 85, 116), GB1 domain (19), its structural homologue protein L (57a), λ-repressor (12), and ribosomal protein L9 (73). In all cases, secondary interactions play no role or only a minor one in determining folding kinetics. Instead, strong and specific longe-range hydrophobic contacts seem to play a dominant role. Numerous experimental observations suggest that DCM is unlikely to provide a realistic description of folding kinetics for small proteins.

Nevertheless, in a recent simulation of a simplified coarse-grained off-lattice model, Zhou & Karplus showed that the DCM scenario can be observed under certain conditions. Specifically, these authors found that under conditions at which helices in isolation are stable, folding of a three-helix bundle Go-model protein may proceed via DCM. Because in most experimentally studied cases isolated helices are hardly marginally stable (33, 34, 84), it remains to be seen whether the DCM observed by Zhou & Karplus can be found in real proteins.

THEORETICAL STUDIES

Experimental results (16, 41, 44, 49, 75, 77), along with a number of computational studies for a variety of models (see below), establish nucleation as a plausible kinetic folding mechanism for small proteins consistent with the cooperative character of their thermodynamics (2, 11, 89, 103, 111). [This does not exclude possible hierarchical mechanisms for higher-order structural organization of proteins such as quaternary structure or multidomain arrangement (49, 76)].

While general validity of the nucleation mechanism of protein folding can be considered established for many small proteins, several crucial details remain unclear. In particular, it is very important to understand what determines the folding nucleus. Protein topology or sequences? This question is closely related to the problem of protein evolution (80, 81). While there are certain indications from simulations and experiments that protein structure may be a strong determinant of nucleus location (2, 15, 17, 75, 80), a complete understanding of what determines the spatial location of the folding nucleus has not yet been reached. Another important question concerns the relative importance of short- versus long-range interactions in nucleation. Recent interesting observations by Plaxco and coworkers concerning a correlation between contact order and folding rate (92) may provide a clue for more in-depth theoretical analysis of this issue. These questions are evolving as central to the field of protein folding (17, 19, 39, 80, 81). Addressing them has been the focus of theoretical studies of folding mechanisms carried out by many groups over the past years.

Determining TSE in Computer Simulations of Simple Models

One of the key issues in computational studies of protein folding kinetics is determination of TSE. Several authors chose to derive TSE from equilibrium (87, 101) or umbrella (113) sampling by determining a one-dimensional free-energy profile

for a certain order parameter. Free energy is defined as:

$$F(A) = -kT \log f(A), \tag{1}$$

where $f(A)$ is the frequency of observing the value of order parameter A in the range $(A, A + \Delta A)$.

The maximum on the plot $F(A)$ at some value of A^* is identified with the transition state. Presumably, TSE consists of conformations with $A = A^*$. This approach was first used by Sali et al to determine the TSE for a lattice 27-mer model (101). Q, a normalized number of native contacts, was chosen as an order parameter for sampling (101). This choice was motivated by earlier analytical studies of random and designed heteropolymers (100, 101) where Q, the overlap with the native state, was shown to be a key order parameter in thermodynamic analysis.

Later Onuchic and coauthors (87, 115) used the same approach of equilibrium sampling and the same order parameter Q to derive TSE for a related 27-mer lattice model. These authors proposed a "multiple delocalized nuclei" model (87) as an alternative to the specific folding nucleus (SFN) mechanism proposed earlier by Abkevich et al (2), who used a different method of search for the TSE. The arguments against SFN (87) were based on the observation that in the ensemble of conformations having $Q = Q^*$ no specific contacts or interactions were clearly dominant.

However, the approach to determine the TSE from equilibrium sampling of any order parameter may be problematic. Du et al directly evaluated the correlation between order parameter Q and transmission coefficient p_{fold} for various lattice models and found no correlation between the two (30). More specifically, the distribution of p_{fold} in the ensemble of conformations with $Q = Q^*$ was found to be very broad, ranging from 0 to values close to 1. This analysis suggests that Q is not a viable RC for protein folding. Further studies by Angerman & Shakhnovich (unpublished results) and by Li et al (69) supported this conclusion, again showing no correlation between Q and p_{fold} (i.e., between Q and TSE) for various models. This analysis clearly shows that it may not be correct to determine the TSE from equilibrium sampling (8, 87, 113). The reason for failure of equilibrium or umbrella sampling methods to provide an adequate low-dimensional description of kinetics, and the TSE was explained in more detail (104).

The TSE for complex systems, such as protein folding models, can be determined only directly from kinetics in the cases when RC is not known (2). An approach to determine TSE from kinetics was proposed (2). This study focused on contacts that appeared on a steep part of folding trajectories (Figure 1) just preceding a folding event in time. Special attention was paid to check explicitly that conformations identified are indeed postcritical (that they feature $p_{fold} = 1$).

A related approach was taken (69) to study the folding nucleus in a more complex model, cubic lattice with side chains (59). Conformations that were close in time to the native state yet had low structural similarity to it ($Q \approx 0.40$) were identified as putative PCCs. Each of these was subjected to the test of running

simulations that started from these conformations. Only a small fraction of putative PCCs turned out to be real (with $p_{fold} > {}^1/_2$), committed to fast folding. This calculation made it possible to identify a folding nucleus in the lattice model with side chains, and interestingly, a few nonnative interactions were found to play an important role in determining the folding nucleus in this model. This finding helped to rationalize some experimental observations with SH3 domains (41, 75).

Another approach to determine TSE in lattice simulations was proposed (42). A virtual protein engineering (PE) study was carried out for a lattice model protein where each residue was mutated to all 19 possible alternatives. Folding and unfolding rates of the mutants were determined and ϕ-values were obtained using a standard definition (49). In comparison with real experiments, the simulations have two major advantages: (*a*) The TSE and the intermediates can be evaluated independently from simulations by other methods (see above) without resorting to the PE analysis. This provides a valuable reference point to evaluate strengths and weaknesses of the PE method. (*b*) Artificial mutations that change the energy of certain particular contacts are possible in simulations. This provides a way to evaluate the degree to which specific (native and nonnative) contacts are formed, without the PE analysis being obscured by the fact that real mutations change a multitude of contacts simultaneously. The results of the study provide detailed guidelines for interpreting PE experiments. The results support the view that the PE approach may be a good way to evaluate TSE provided that multiple mutation scans are made. On the other hand, the analysis (42) pointed to certain limitations of the PE method. The most important of them is that ϕ-values get unreliable when mutations result in small changes in stability, $\Delta\Delta G$. The main reason is not an obvious increase in error bars when the denominator gets smaller, but rather a possible peculiar compensation effect of competing strong interactions of various magnitude and sign (see details in 42).

Furthermore, the analysis (42) addressed the issue of the temperature dependence of the TSE. This analysis provided, for the first time, a microscopic interpretation of considerable temperature shifts in ϕ-values that were observed in CI2 by Fersht and coworkers (86).

Perhaps the most direct way of determining the TSE was suggested by Du et al (30) and later used (24, 24a, 88, 89a, 90). [A similar method was introduced by Karplus and coworkers to find the transition state for activated processes in proteins (85a).] The idea of the method is to find a set of conformations that have $p_{fold} \approx {}^1/_2$. This is achieved by starting simulations from numerous conformations obtained from folding or unfolding trajectories. The main disadvantage of the method is its extreme computational intensity. Using this method, Dinner & Karplus (24a) studied folding of 125-mer on the cubic lattice. In accordance with earlier observations (79), they found the fast and the slow track trajectories. Slow folding is attributed to the trapping in the off-pathway misfolded conformations. The trapping occurs before the TSE is reached.

Several kinetic methods of analysis applied to lattice models of various degrees of complexity provided a consistent view on the character of PCCs that share a

common folding nucleus. However, this conclusion was questioned by Klimov & Thirumalai (60) who studied similar lattice models but failed to detect a specific nucleus. The reason for this discrepancy was explained (105): Klimov & Thirumalai collected and analyzed conformations that appear prior in time to reaching the native state (i.e., they collected putative PCCs). These authors failed to check which of the putative PCCs are actual postcritical ones (with $p_{fold} > {}^1/_2$). This test is a crucial part of the search for PCCs. Only a small fraction (less than 20%) of putative PCCs are actual ones (committed to fast folding) (69). Thus, the analysis of Klimov & Thirumalai has a technical problem that prevented them from correct determination of the actual set of PCCs.

The kinetic analysis of simulations, when properly applied to lattice model MC simulations, points to a specific nucleus as the defining feature of the PCCs.

Nucleation in Off-Lattice Models

How universal is this conclusion? Is it transferable to a more realistic, off-lattice model and/or another dynamic simulation algorithm? This question was addressed in a recent publication (28) where off-lattice folding was simulated using discrete molecular dynamics (27, 122). The authors used a Go model (116a) to study folding of a small (56 residues) off-lattice protein. The thermodynamics of this model was presented in detail in an earlier publication (27) where the folding transition was cooperative. The search for a folding nucleus in the off-lattice model (28) consisted of the analysis of conformations obtained in deep equilibrium folding and unfolding fluctuations. The idea of this analysis is that partly unfolded conformations that are committed to immediately returning back to the folded state are the conformations that retain a folding nucleus, while fluctuations from partly folded conformations that return back to an unfolded state represent conformations that have not formed the folding nucleus. Comparison of such "folded-folded" fluctuations with "unfolded-unfolded" ones made it possible to identify the folding nucleus in the off-lattice Go-model of a protein. Further testing showed that fixation of nucleus contacts eliminated the barrier between folded and unfolded conformations, whereas fixation of the same number of control nonnucleus contacts did not change the landscape qualitatively and retained the barrier (Figure 2).

Topology May Define the Folding Nucleus: Key Findings from Simulations and Experiment and Its Evolutionary Implications

The results of lattice and simple off-lattice simulations resulted in a remarkable conclusion that the location of the folding nucleus may be determined to a greater extent by the native structure than by the details of sequence that fold into that structure. This discovery was made in 1994 by Abkevich et al (2). It was found that the folding nucleus was identical for 30 sequences designed to fold into the same lattice conformation, despite the fact that sequences were quite different. It was concluded that the nucleus location is determined primarily by the native

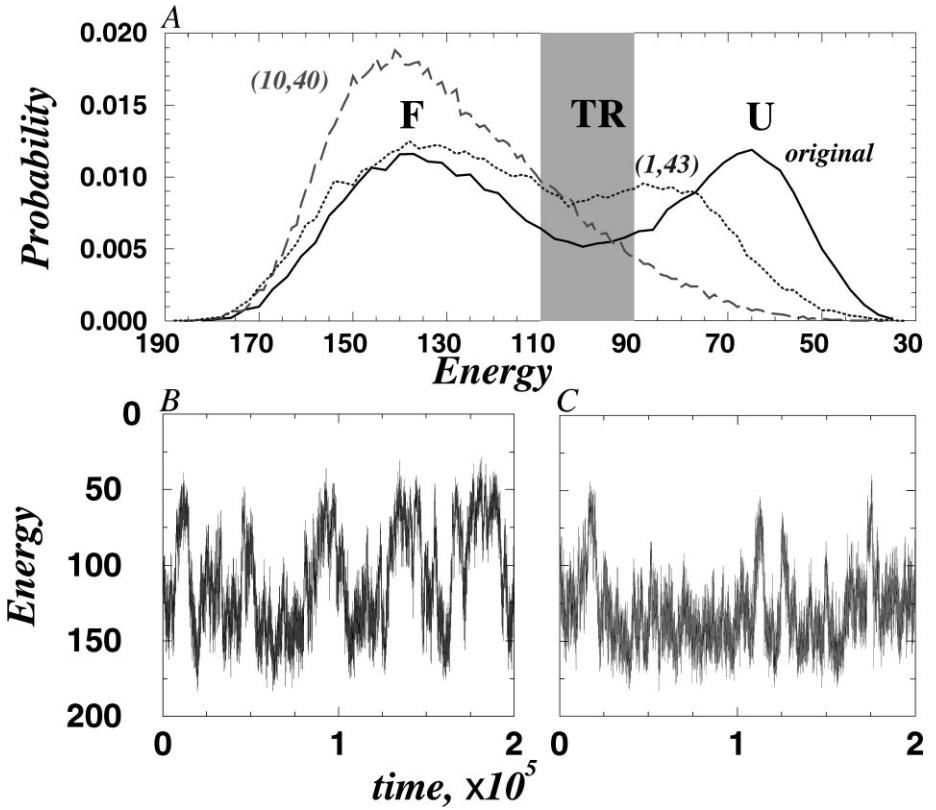

Figure 2 Specific nucleus mechanism in the off-lattice model. Permanent fixation of a nucleus contact 10–40 eliminated the barrier between the folded and unfolded state (*A*, dashed curve), while fixation of a control nonnucleus contact did not change the character of the folding transition (*A*, dotted curve). Below are equilibrium folding trajectories for the original wild-type chain (*B*) and one with a fixed nucleus contact (*C*). TR denotes the transition region of energies where the TSE belongs.

conformation (2). This conclusion was further supported by lattice-model studies where two 48-mer sequences were designed, using different potentials, to fold to the same conformation. These sequences were found to have identical folding nuclei (107) (despite the fact that different amino acids were placed in the nucleus locations for different potentials). Other evidence in favor of the primary role of structure in determining the folding nucleus comes from the simulations of the off-lattice Go-model (28) where all native interactions are of the same magnitude and yet a specific nucleus was found.

The evolutionary implications of the primacy of structure over sequence in determining the folding nucleus were realized (80) where it was pointed out that each structure may feature special nucleation positions that serve as "accelerator

pedals" for folding into that structure. The folding rate then can be controlled by placing specific amino acids into those positions that make the nucleus stronger or weaker depending on the rate requirements. In a dramatic demonstration of this principle, sequences evolved under selection pressure to fold fast. The simulated evolution protocol consisted of random mutations that were accepted if mutant sequences folded faster and were rejected otherwise. One thousand independent folding runs were made if a mutation was accepted in order to obtain an unbiased estimate of the current wild-type folding rate. A slight variance between estimates of folding rate for various runs provided noise to the algorithm, resulting in some instances when slightly slower folding mutants were accepted. This is analogous to temperature in more traditional Monte-Carlo techniques.

The selection algorithm provided a large number of fast-folding sequences for lattice model proteins. Statistical analysis of them showed that the selection pressure was applied primarily at the nucleus position for that structure. In particular, nucleus contacts were enforced by amino acids that strongly attracted each other. It is interesting to note that in another run of the evolutionary algorithm for the same structure, another family of fast-folding sequences was generated that had the same striking feature of an enforced nucleus. The nucleus for the new family was identical to that for the previously generated family, although the types of amino acids placed there were different. For both independently generated families, the nucleus positions were among the most conserved ones in intrafamily alignment.

The evolutionary implications of these findings were explored and further analyzed (80, 81, 87a; see below).

Experimental results demonstrate that proteins sharing the same fold and a very low sequence similarity have similar structure of the transition state (15, 17, 41, 44, 75, 75a, 90a, 100a, 121). In other words, the same secondary structure elements are involved in the formation of the folding nucleus. Furthermore, changing the connectivity of a protein through circular permutations modifies the TSE (120).

On the other hand, mutations in the folding nucleus can accelerate protein folding by almost two orders of magnitude without affecting protein topology (37). There are also examples of proteins having similar folds in which the folding nucleus is found experimentally in different regions of the protein [U1A, S6 and Ada2h (117, 121), protein G and protein L (57a, 76a), SH3 domain and Ssso7d (41a)].

The key conclusion made from lattice model studies, that native structure may determine the folding nucleus (2, 80, 107), inspired a series of experimental works seeking to test/verify this conclusion (15, 44, 121) and a number of computational approaches attempting to predict the TSE and the folding nucleus from the native structure for several real proteins (4, 5, 18, 39, 83, 94, 114).

The role of the native structure in determining essential features of the folding kinetics was further highlighted by the observation of correlations between folding rates and certain structural properties of native proteins (92).

Contact Order: What Really Matters?

As more proteins were studied, a trend became apparent: On average, helical proteins fold faster ($k_f^{H2O} \approx 10^2-10^5\ s^{-1}$) than most β and α/β proteins ($k_f^{H2O} \approx 10^{-1}-10^3\ s^{-1}$). Clearly, local interactions can be formed fast and can even be present in the unfolded state under native conditions (95). Hence, it is natural to assume that higher content of local interactions (as in the helical proteins) leads to faster folding.

On the other hand, several proteins of the same fold topology were shown to fold with very different folding rates. The main examples here are proteins of the immunoglobulin domain ($k_f^{H2O} = 1.5-155\ s^{-1}$), SH3 domain ($k_f^{H2O} = 0.35-94\ s^{-1}$), cytochrome c ($k_f^{H2O} = 400-15000\ s^{-1}$), and proteins of α/β-plaitfold ($k_f^{H2O} = 0.23-897\ s^{-1}$) (50). It is not clear to what extent topology of the native structure alone can explain these data.

Plaxco et al (92) suggested that the average distance between residues interacting in the native state (contact order) can be used as a general descriptor of protein topology to correlate topology with the folding rate. Contact order is defined as

$$CO = \frac{1}{NL} \sum_{i,j} \sigma_{ij} \, |i - j| \,, \qquad (2)$$

where $\delta_{ij} = 1$ if residues i and j are in contact and 0 otherwise; N is the total number of contacts, and L is the protein length. For a number of two-state folding proteins, contact order was reported to exhibit a statistically significant correlation with the log of the folding rate in water ($\log k_f^{H2O}$) (50, 83, 92).

In a recent study, Dinner & Karplus (23a) focused on the role of stability and topology of the native state in determining the rates of folding. They trained neural networks to reproduce k_f of 33 proteins from their CO and ΔG. In contrast to (92), they carefully cross-validated the results by leaving aside each group of structurally related proteins while training the network. Prediction of k_f using CO alone gives correlation $r = 0.73$. Stability $\Delta G/N$ and CO taken together yield $r = 0.79$, whereas stability alone gives $r = 0.37$. While emphasizing the role of stability in determining the folding rate, these results support topology as the main determinant to the folding rate.

Why Does Contact Order Correlate with Folding Rate?

The contact order is clearly related to secondary structure: α-helical proteins have large numbers of local contacts and hence have low contact order. In contrast, β and α/β proteins have many distant or nonlocal contacts and hence have greater contact order. Because helical proteins (on average) fold faster than β and α/β proteins, an obvious anticorrelation between folding rate and contact order emerges. However, contact order was suggested to capture more of protein topology than just secondary structure.

We decided to check whether contributions other than helical content are important for correlation between CO and the folding rate. We express contact

order as

$$CO = \frac{1}{L}[\, fL_{local} - (1 - f)L_{distant}],$$

where $f = N_{local}/N$ is a fraction of local contacts, i.e., contacts between residues i and j, such that $|i-j| \leq 4$; $L_{local} = (\sum_{|i-j| \leq 4} \delta_{ij}|i-j|)/N_{local}$ is the average separation between residues forming local contacts; and $L_{distant}$ is the average separation between residues forming distant ($|i-j| > 4$) contacts. This way of presenting *CO* separates terms that come from the helix/turn content (f) and those that take into account the actual distribution of local and distant contacts in the native structure (L_{local} or $L_{distant}$).

The question is which of the three components, f, L_{local}, or $L_{distant}$, correlates with $\log k_f^{H2O}$. For this test we have chosen a challenging family of proteins, all having the same α/β-plait fold, but no evident sequence similarity. Importantly, the six studied proteins span folding rates from 0.23 s^{-1} to 897 s^{-1}.

Figure 3 presents correlations between $\log k_f^{H2O}$ and each of *CO*, f, L_{local}, and $L_{distant}$ as a function of contact cutoff distance, R_c. (For *CO*, L_{local}, and $L_{distant}$ we changed the sign of the correlation coefficient.) First, notice that *CO* strongly depends on the definition of contact R_c exhibiting low correlation for $R_c < 6.5$ Å and $R_c > 9$ Å. Second, the fraction of local contacts f (i.e., the measure of helix/turn content) exhibits correlations higher than *CO* over all values of R_c except 7 Å $\leq R_c \leq$ 7.5 Å, where both correlations are high (above 0.9) and close. Third, L_{local} and $L_{distant}$ have no significant correlation with $\log k_f^{H2O}$. From this example, we conclude that it is the content of local contacts that makes *CO* correlate with the folding rate, not the fine details of contact distribution, sequence separation of distant contacting residues, etc. In Equation 2, one can substitute $|i-j|$ with 1 for all local contacts and with 0 for all distant ones and get greater correlation with $\log k_f^{H2O}$ for most of R_c values. Figure 4 shows how contact order and fraction of local contacts predict folding rates for α/β-plait proteins at 7 Å cutoff.

Detailed distribution of loop lengths, however, is important when the folding rate of loop-insertion mutants is discussed. Recently, Fersht suggested a model that links contact order to the change in folding rate upon loop-insertion mutation (37). He showed that when a loop of n residues is extended by l residues, the change in the entropy of the TSE is given by $\Delta\Delta S \sim \log(1 + l/n) \sim l/n$ for $l \ll n$. This, in turn, leads to the increase in the activation free energy, $\Delta\Delta G^{\ddagger} = \Delta\Delta E - T\Delta\Delta S \sim -l$, and a decrease of folding rate, $\Delta \log k_f = \Delta\Delta G^{\ddagger} \sim -l$. On the other hand, loop extension leads to linear increase of the contact order ($\Delta CO \sim l$) and hence $\Delta \log k_f \sim -\Delta CO$. Experimental data for CI2 and SH3 fit this linear model very well, with the slope of the same order of magnitude as predicted by Fersht's model.

Fersht also emphasized the importance of specific interactions in the folding nucleus as primary determinants of folding rate. Clearly these interactions are missing from the contact order, which takes into account only the entropic term. For example, mutations in the folding nucleus of CI2 do not change the contact

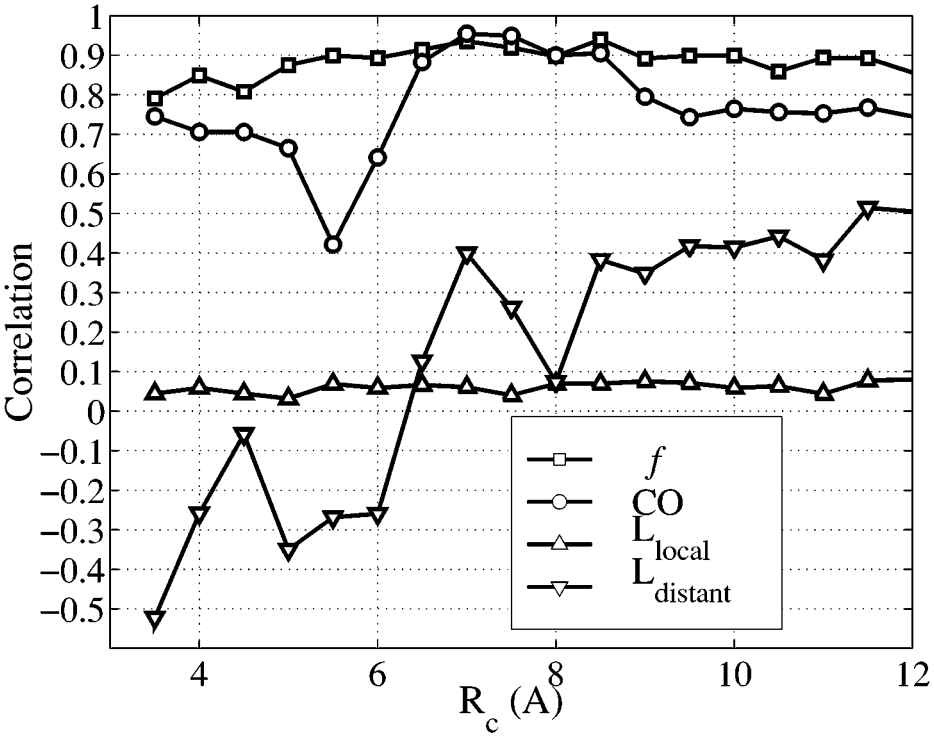

Figure 3 Correlation between the folding rates $\log k_f^{H2O}$ and the contact order (CO), the fraction of local contacts (f), the average length of the local (L_{local}) and distant ($L_{distant}$) interactions as a function of the contact cutoff R_c (see text for details). The correlation for six α/β-plait fold proteins. All correlations are expressed as positive for easy comparison. PDB files a in Table 1. Summation is over all contacts such that $i \leq j + 2$.

order much (if at all), but result in a three order of magnitude change of the folding rate, from 2.4 s^{-1} for the double mutant A16GI57A to 2300 s^{-1} for R48F. Jackson (50), in her review of single-domain proteins, presents an example of ROP protein, which upon mutations in the hydrophobic core, starts folding (and unfolding) three orders of magnitude faster. Several other examples clearly demonstrate that topology is not the sole determinant of the folding rate.

Simple Models: Topology-Dependent Free Energy Functionals

Munoz & Eaton suggested a simple statistical model to predict folding rates and ϕ-values for two-state proteins (83). In their model, each residue can be in one of two states (native or denatured). In the native state, a residue loses Δs of entropy (the value depends on its secondary structure) and gains Δe of energy from interactions with other residues that are in the native state. The energy of the interactions

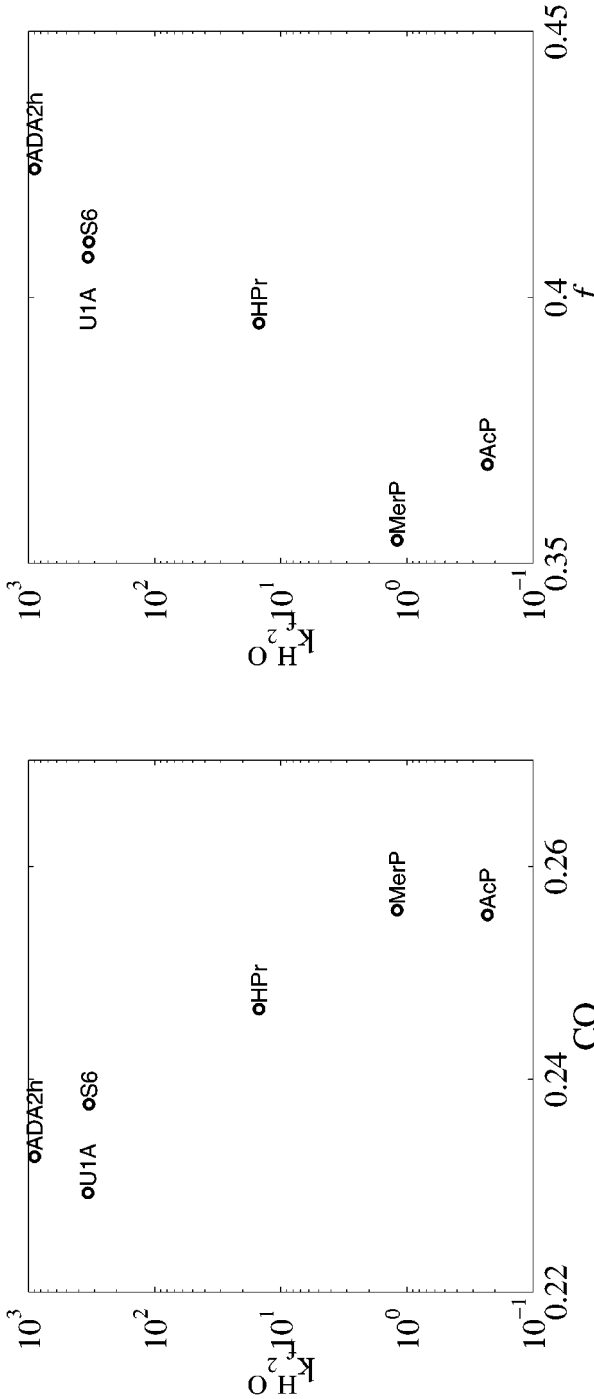

Figure 4 The folding rate $\log k_f^{H2O}$ versus CO (left) and versus f (right) at a contact cutoff $R_c = 7$ Å.

TABLE 1 Folding nuclei as identified by the authors

Protein	PDB	Folding Nucleus	Reference
CI2	2ci2I	A35 L68 I76	49
Tenascin	1ten	I821 Y837 I860 V871	44
CD2.d1	1hnf	L19 I21 I33 A45 V83 L94 W35	72
CheY	3chy	D12 D13 D57 V10 V11 V33 A36 D38 A42 V54	71
ADA2h	1aye	I15 L26 F67 V54 I23	121
AcP	1aps, 2acy	Y11 P54 F94 Y25 A30 G45 V47	15
U1A	1urn	I43 V45 L30 F34 I40 I14 L17 L26	117
ACBP	1aca	F5 A9 V12 L15 Y73 I74 V77 L80	62
FKBP12	1fkj	V2 V4 V24 V63 I76 I101 L50	74

between two residues depends on the number of contacting atoms. In order to reduce the number of configurations in the model, only one or two regions of the native structure are permitted simultaneously in a protein. Each of these regions is a continuous part of the chain where all residues are in the native state. A crucial simplification is that a residue interacts only with other residues in its native region and not with residues of another region. Basically, the native regions are independent and, hence, the long-range interactions are downplayed. Dependence of the entropy of a loop on the loop length is also ignored. Despite these simplifications, predicted folding rates correlate well ($r = 0.87$) with the experimental ones. This is an improvement compared to simple contact order predictions (see above). This improvement, however, comes at a cost of many adjustable parameters used in the model: two values of Δs, the diffusion coefficient D, and most importantly, the energy of an atomic contact ε that is chosen separately for each protein. The authors generated a set of Δss and εs that are consistent with protein thermal stability, but then chose values from this set that minimized the squared residuals between the observed and calculated folding rates for the 18 two-state proteins. Adjustment of 18 εs to fit folding rates for every protein severely diminishes the predictive power of the method. It is unclear how well the method will perform without adjustment of parameters to maximize correlation with the data the method seeks to predict. Computed ϕ-values for CI2 have very low correlation with experimental ones (correlation coefficient not reported). The use of noninteracting native regions eliminates the possibility of a transition state with natively interacting residues separated by denatured loops.

A related model was developed by Alm & Baker (4). Similar to the model of Munoz & Eaton (83), each residue can be in either of the two states, and only the conformations with one or two stretches of the native residues are considered. The free energy of a conformation is computed differently. It has three terms: (*a*) The attractive interactions of the native residues are proportional to their buried

area (both hydrophobic and polar); (*b*) the entropic cost of ordering a residue is proportional to the number of native residues; (*c*) the entropy of the loop between the two native stretches is proportional to the log of the loop length. Importantly, interactions between the two stretches are still ignored; that is, each stretch is contributing independently, and its contribution to the energy does not depend on the length or amino acids in the other stretch. Kinetics is modeled by a series of steps of extension/shrinking of the stretches by one residue. This allowed full enumeration of the states and identification of the lowest free energy path from the fully unfolded to the fully folded state. Transition states (peaks on the free energy) were obtained from the low energy paths and used to compute the folding rates and ϕ-values. Parameters of the model were adjusted to maximize correlation with the experimentally observed ϕ-values of CI2, SH3, and barnase. The model fails to predict the folding rates better than the contact order. Predicted ϕ-values, however, show good correlation $r = 0.5.0.87$ for five out of seven studied proteins (two of these five were not used to adjust parameters). The ϕ-value prediction fails for procarboxypeptidase activation domain and for L protein. Unfortunately, only correlation coefficients were presented, and detailed predictions for the studied proteins were not reported (4), making it difficult to evaluate which aspects of the transition state were correctly predicted by the model.

The model developed by Galzitskaya & Finkelstein (39) combines detailed atom-atom treatment of the interaction energy with the loop entropy computed in a way similar to the calculations of Alm & Baker (4). Up to four stretches of interacting native residues are allowed in this model. Importantly, in contrast to models described above, Galzitskaya & Finkelstein explicitly consider interactions between the individual residues belonging to different stretches. This model has only one adjustable parameter. A dynamic programming technique is used to find transition state conformations, defined as peaks on the lowest energy path from the unfolded to folded state. Although approximate, dynamic programming produces a reasonably good correlation between computed and experimentally measured ϕ-values for three of five studied proteins (CI2 $r = 0.56$, barnase $r = 0.54$, CheY $r = 0.50$, α-spectrin SH3 0.39 with only six experimental points). ϕ-values obtained for src SH3 do not correlate with experimental ones. The use of dynamic programming became possible at the expense of a strong unphysical assumption that local unfolding is excluded on each pathway.

Each of these three models involved strong assumptions that were difficult to justify on physical grounds. Probably the most striking one is the assumption that each amino acid can be in two states, native and denatured, and the ability to be in the native state is independent of the state of all other residues. Such an assumption is normal for one-dimensional systems but may be inappropriate for three-dimensional proteins because the native state of a residue depends on its contact with structural neighbors. The folding process is assumed in these models to propagate via native interactions, and distant parts of the protein can interact in the models of Munoz & Eaton (83) and Alm & Baker (4) only when all residues between them are native, i.e., the configurations where long loops connect

interacting parts (see Figure 4 of reference 49) are excluded. Consecutive residues are grouped into effective ones and local unfolding is excluded (39). Importantly, the Munoz-Eaton and Alm-Baker models are fundamentally one-dimensional. Therefore, they cannot reproduce the folding phase transition, and hence, they may not be fully adequate for description of the folding energy landscape.

Success of these models may be due to the fact that they implicitly explore some simple features of the native structure that may be correlated to some extent with ϕ-values. This was addressed in a recent work (N Dokholyan, L Li, E Shakhnovich, manuscript submitted) where correlation between ϕ-values and the number of contacts that an amino acid makes in the native structure was studied. Pronounced correlation was observed in 6 of the 11 proteins studied, making this very simple analysis comparable in performance with the models studied (4, 39, 83). This is even more striking given that the analysis (N Dokholyan, L Li, E Shakhnovich, manuscript submitted) did not involve any adjustable parameters or assumptions in contrast with other studies. However, it is important to note that the success of all oversimplified models (4, 39, 83; N Dokholyan, L Li, E Shakhnovich, manuscript submitted) can be considered as only moderate, making the need for a more consistent physical theory of folding kinetics, which is free of uncontrollable assumptions, even more pressing.

Shoemaker and coauthors (114) proposed a more rigorous approach based on presenting a free energy of the protein chain as a phenomenological functional of all formed native contacts $\{Q_{ij}\}$. The functional contains many empirical terms and parameters such as contact energies as well as many additional *ad hoc* cooperativity terms introduced to account for several cooperativity effects caused by polymer connectivity. The entropy contribution is in the form of a combination of Jacobson-Stockmayer loop entropy terms, taken independently for each contact, with the additional contribution corresponding to entropy reduction due to formation of μ quenched contacts. Additionally, an entropic contribution of a combinatorial form, yet related to each individual contact, is included. The TSE was assumed to consist of all conformations having a given number Q^* of native contacts; the value of Q^* was taken from experiment (114). The authors evaluated the probability of each contact by minimizing their free-energy functional with respect to each individual contact under the additional condition that the total number of native contacts equals Q^*. A number of parameters were adjusted in the model, and a fair agreement between the predictions for TSE and experiment was reported.

While such an agreement is encouraging, it is difficult to say what is the main physical reason for it: The free-energy functional featured many terms that do not follow from microscopic analysis, and it is not quite clear which ones are responsible for the observed results.

A related approach was taken by Clementi and coauthors (18), who also assumed that all conformations with a specific number of native contacts corresponding to the maximum of the curve $F(Q)$ of free energy versus Q constitute the TSE. These authors simulated the off-lattice simplified models of several proteins using

Go-type energetics that biased both interaction energies and bond and dihedral angles toward their native values [the latter contributions from angles were necessary to get a maximum on the $F(Q)$ curve obtained in equilibrium sampling using a multiple-histogram method]. Obtained TSE conformations were reported to be in reasonable agreement with experimental ϕ-values.

Another interesting approach to determine kinetically hot residues was developed by Demirel et al (22), who used the Gaussian network model (GNM). The basic idea of the GNM is to consider a protein structure as an elastic network of harmonic interactions between contacting residues. The frequency of fluctuations in the network is computed for each residue. Residues with high frequency are called kinetically hot. Demirel et al report that kinetically hot residues in CI2, Cytochrome C, and CheY correspond to positions "critically important for the stabilization and folding process." It is unclear whether the GNM is able to separate residues important for stability from those crucial for folding kinetics. Another question is whether the GNM frequency of a residue is more informative than a simple solvent accessibility. Although elegant and physically motivated, the GNM predictions of kinetically hot residues need to be compared with more up-to-date ϕ-value measurements.

EVOLUTIONARY STUDIES

A Successful Blind Prediction of Folding Nucleus from Conservation

A first successful prediction of the nucleus residues from protein structure was made (107), where many sequences were designed to fit the structure of CI2 with low energy. Positions conserved among the designed sequences were identified as the putative nucleus, and blind predictions were made as to which residues belong to the folding nucleus. Remarkably, subsequent experiments (49) independently corroborated the theoretical predictions. It was pointed out (107) that the method was successful probably because the sequence design procedure was able to identify a contiguous tight cluster of strongly interacting residues and conservatively placed the most strongly interacting amino acid types there. Interestingly, comparison of the design entropy with conservation in a family of aligned real sequences homologous to CI2 (taken from the HSSP database; 26) revealed remarkable conservation in the nucleus positions as predicted from sequence design simulations (107). Clearly, it was not appropriate to consider just a correlation coefficient between predicted and observed conservatism; in real sequences, positions can be conserved for many biological or historical reasons unrelated to the nucleation mechanism or even to structural factors (active site conservation). Nevertheless, the positions where design entropy is consistent with conservatism derived from sequence alignment may point to positions that are conserved for structural, rather than functional, reasons because sequence design is sensitive only to structural energetic factors. It is interesting to note that for several conserved nucleus positions, the design simulations were able to correctly predict amino acid types, suggesting that real protein

sequences may indeed have been optimized for stability (that is, the optimization strategy adopted in design). Furthermore, the agreement between natural sequences and designed ones in conserved positions suggested that the energy function used in these design simulations was meaningful, despite the simplicity of approximation (contact interactions between C_β atoms at a contact cutoff distance of 7.5 Å).

Success in prediction of the folding nucleus from conservatism analysis motivated researchers to look at other proteins. In a series of papers, Ptitsyn carried out a detailed analysis of conservatism in distant yet related by sequence homology members of the cytochrome C (97) and myoglobin (99) families. In both cases, highly conserved clusters of residues without an obvious functional role were found. It was suggested that the amino acid residues forming those clusters constitute folding nuclei for their respective proteins. Michnick & Shakhnovich (78) carried out an analysis of conservation for families of three structurally related proteins—ubiquitin, raf, and ferredoxin.

At sites where there was a correlation of low entropies between designed and natural sequences, additional tests were applied: In order to qualify as a putative nucleus, amino acids at these positions had to be conserved both in ubiquitin and raf families and make contact with at least one of the other conserved sites (to form a nucleating cluster). Such stringent criteria made it possible to identify seven potential nucleation sites in ubiquitin and raf. Most of them are aliphatic hydrophobic amino acids in both proteins. However, importantly, not all hydrophobic core residues were identified as nucleation sites in this analysis (e.g. 7 out of 16 sites in raf and ubiquitin). While there is no ϕ-value analysis of ubiquitin folding transition state, the predicted nucleus postitions are equivalent to positions in structurally homologous protein L that were found to be conserved in fast-folding sequences that emerged in evolution-like phage-selection experiments (57).

Folding Nucleus, ϕ-values, and Evolutionary Conservation

A recent paper by Plaxco et al (91) addresses the question of whether high ϕ-value residues are more conserved than low ϕ-value residues. By comparing conservation of low and high ϕ-value residues in six proteins, they came to a conclusion that in most of the cases high ϕ-value residues are not more conserved than the residues that have low ϕ-values. This result comes as no surprise. However, this almost obvious observation (see below) does not justify the conclusion made by the authors that the *folding nucleus* is not more conserved than the rest of the protein. Why do we expect no straightforward correlation between the ϕ-values and the conservation?

First, buried (mostly hydrophobic) residues are known to be more conserved than the rest of the protein, on average. Many of those buried amino acids are irrelevant for kinetics: They are conserved to provide protein stability and hence exhibit low ϕ-values. Each family of homologous proteins also has several residues conserved for functional reasons (active/binding site) or historic reasons (not enough evolutionary time to diverge). Hence, one would expect any ϕ-values for highly conserved residues. Second, low ϕ-value is not a very good reference set, as PE

analysis usually focuses on somewhat buried residues, which in turn, tend to be more conserved. For some proteins (e.g. ACBP), experimental groups intentionally focused their analysis on conserved residues. Third, there is no one-to-one correspondence between the ϕ-value and the participation in the folding nucleus: Neither a high ϕ-value implies greater contribution to stability of the transition state, nor do all residues in the nucleus have high ϕ-values (44, 49, 74, 117). Several residues that have high and moderate ϕ are drugged into the transition state but are not very important for the stabilization of either transition or native states. That is, residues on the exposed side of the sheet in FKBP12 (74) diverging turn in SH3 (120), tenascin (44). These arguments show that similar levels of conservation in the low and high ϕ-value residues do not prove or indicate that the folding nucleus is no more conserved than the rest of the protein.

Is Folding Nucleus Conserved?

Why do we expect the folding nucleus to be more conserved than the rest of the protein? First, residues in the folding nucleus are conserved because they are important for stability of the native structure! In fact, in most of the studied proteins, the nucleus is a dense cluster of residues stabilized by either hydrophobic or hydrogen bonding interactions. Usually this cluster is part of the protein core that is more conserved because it provides stability to the native structure. Second, residues in the nucleus are conserved because they provide fast folding. If slow folding affects normal protein function (leads to aggregation, early proteolysis, etc) and, hence, results in a disadvantageous phenotype, then amino acids in the nucleus must be under additional evolutionary pressure. Even if the requirement to preserve protein folding rate is much weaker than the requirement to preserve protein stability, we expect the folding nucleus to be noticeably more conserved than the rest of the protein.

In order to check whether the folding nucleus is really conserved and to incorporate data from a few recent experiments, we performed an extensive study of conservation patterns in nine proteins: CI2, FKBP12, ACBP, CheY, Tenascin, CD2.d1, U1A, AcP, and ADA2h (L Mirny, E Shakhnovich, manuscript submitted). Five of them are the same as studied by Plaxco et al (91).

In contrast to Plaxco et al, we (*a*) defined folding nucleus as it was identified by the original experimental groups (Table 1), (*b*) grouped residues into classes to reflect the natural pattern of substitutions, (*c*) compared conservation of the folding nucleus with the conservation of all residues in the protein, and (*d*) computed conservation (sequence entropy) as

$$s(l) = -\sum_{i=1}^{6} p_i(l) \log p_i(l), \tag{3}$$

where $p_i(l)$ is the frequency of residues from class i in position l of multiple sequence alignment.

Figure 5 presents a conservation profile for studied proteins with nucleation positions marked by filled circles. In contrast to the claim made by Plaxco and

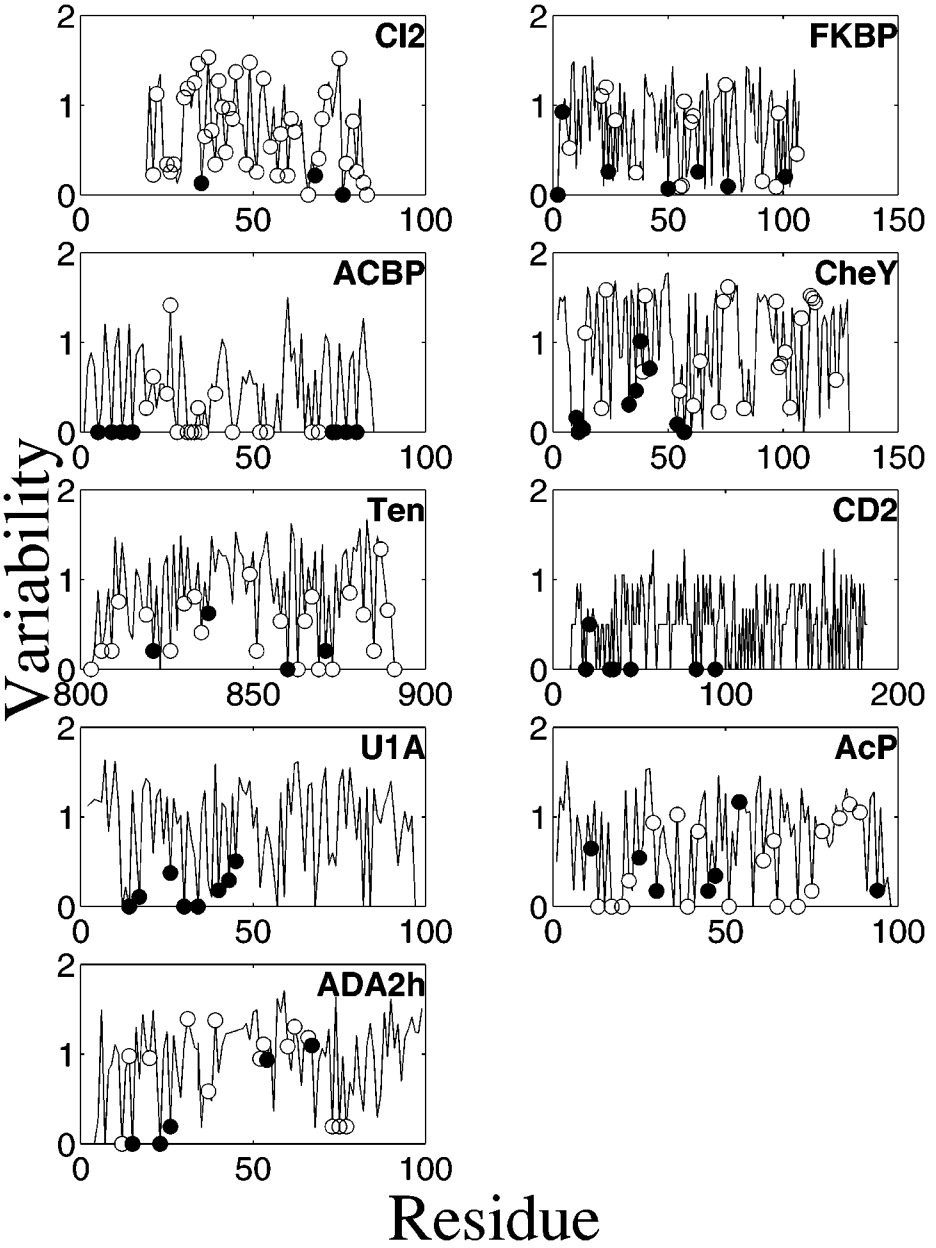

Figure 5 Variability profiles (sequence entropy) for nine different proteins computed using six types: [AVLIMC][STNQ][KR][DE][GP][FWYH] (80a). Circles indicate positions at which ϕ-values have been experimentally measured. Residues forming the folding nucleus are shown by filled circles.

coauthors, for all proteins, except AcP, residues in the folding nucleus are considerably more conserved than the rest of the protein. Statistically, the probability that the observed stronger conservation in the nucleus positions just by chance is 0.0041 (CI2), 0.0187 (FKBP), $<10^{-5}$ (ACBP), $<10^{-5}$ (CheY), 0.008 (Ten), $<10^{-5}$ (CD2), 0.0009 (U1A), 0.089 (AcP), and 0.0023 (Ada2H). See L Mirny & E Shakhnovich (manuscript submitted) for details and analysis of how the form of Equation 3 and the grouping of amino acids into classes influences the results.

Although, on average, the nucleus is more conserved, not all nucleating residues are strongly conserved. For example, in CheY, two out of ten nucleation residues are not conserved. In ADA2h two out of five and in tenascin one out of four residues are not conserved. Although most nucleation residues in studied proteins (except AcP) are very conserved, some may not be either because they constitute an extended nucleus or because of limitations of our residue classification scheme that puts aromatic and aliphatic residues into two different groups, whereas aromatic-aliphatic substitutions are relatively frequent in the core of some proteins (tenascin, ADA2h). Another interesting observation is that the only protein that exhibits no preferential conservation of the folding nucleus is AcP, which is the slowest folding protein among all studied two-state folding proteins ($k_f^{H2O} = 0.23\,s^{-1}$). Perhaps this protein did not experience evolutionary selection for faster folding, and hence, its folding nucleus is under no additional pressure to be conserved.

Note that, as expected, several other residues in studied proteins are as conserved as the nucleating ones. Those include residues of the active site, core hydrophobic residues responsible for stabilization of the native structure, and others. This indicates that although the folding nucleus is conserved, it cannot be uniquely identified by the analysis of a single protein family, as a pattern of conservation is dominated by residues conserved for protein stability and function (see 81 and next section). Thus, a consistent analysis should discriminate between residues that are conserved for functional, stability, and kinetic reasons (folding nucleus).

Universally Conserved Residues—Conservatism of Conservatism

Recently we suggested a method to find universally conserved positions in conservation patterns of several protein families aligned together (81). The idea of the method is to analyze conservation profiles of protein families that belong to analogous proteins (proteins that share the same fold but have no evident sequence similarity and evolutionary relationship). For each family m we computed a conservation profile $s^m(l)$ using Equation 3, and then averaged profiles over M analogous families that were structurally aligned to each other,

$$S(l) = \frac{1}{M}\sum_{m=1}^{M} s^m(l). \qquad (4)$$

Because analogous proteins have unrelated functions, functional residues are usually located at different positions in the structure (except for the case of

"supersite," see below). Then conserved functional residues of one protein are matched in the structural alignment with nonfunctional residues of the others, and conservation that originates from functional constraints is averaged out. Historic conservation (positions conserved because proteins in the family did not have enough evolutionary time to diverge) is also averaged out because different analogous families have unrelated evolutionary history. We showed that 80–90% of variations in the average profile $S(l)$ are explained by conservation driven by the stability of the native structure (see Figure 1 of 81). In each studied fold, a few positions were much more conserved than expected from stability considerations alone. Those positions are apparently under some additional evolutionary pressure and are conserved in most of the proteins of this fold. Note that different families can place different amino acid types in each of those universally conserved positions. What makes these positions special is that the intrafamily conservation itself is conserved among the families. These positions are said to exhibit conservation-of-conservation (CoC).

This analysis for five major folds [immunoglobulin domain, OB-fold (β-barrel), Rossmanfold, α/β-plait, and TIM barrel] (81) revealed that each of them indeed featured positions with statistically significant CoC. Further analysis identified those positions as related to a functional supersite (TIM-barrel and Rossman fold), the folding nucleus (Ig fold, α/β-plait, and OB fold), or both (Rossman fold). The folding nucleus origin of CoC was confirmed experimentally for the Ig fold (44) and the α/β-plait (15, 121).

The analysis also shows that folding rate selection, while noticeable, is relatively weak in comparison to selection on protein stability because most of the signal in the mean conservation profile is explained by stability. However statistically significant, CoC indicates the presence of selection on the rate of folding.

Perhaps the most convincing evidence for evolutionary selection of folding rates comes from the analysis of SH3 domains. Serrano and coworkers (75) studied ϕ-values for the α-spectrin SH3 domain, and Baker and coworkers carried out a similar ϕ-value analysis for src SH3 domain (41). Both authors found a position in the structure that exhibited an anomalous ϕ-value. I34 in src SH3 (and its structural analog in spectrin SH3) seemed to form strong nonnative interactions in the folding nucleus judged by its anomalous ϕ-value (3.9). I34 appears to be a kinetically, but not thermodynamically, important residue in src SH3. Simulations of the lattice model with side chains supported this interpretation, pointing out the importance of nonnative interactions in the folding nucleus (69). Quite remarkably, I34 appeared to be under strongest evolutionary pressure as revealed by the CoC analysis (Figure 6; 69). This observation constitutes the most direct evidence that evolutionary pressure controls the folding rate, as well as stability or function.

Recently Ortiz & Skolnick suggested using correlated mutations to predict kinetic "hot spots" (87a). The authors studied lattice proteins that were evolved to fold fast and for which folding nucleus was identified (80). Ortiz & Skolnick found that correlated mutations between positions around the folding nucleus arise in a statistically significant manner. Qualitatively similar results were obtained for

real proteins. Unfortunately, authors neither list hot-spot positions they found nor provide correlations of predicted and observed ϕ-values.

So CoC is still a method of choice when enough data are available because it can provide very specific indications of the folding nucleus. Unfortunately, the CoC analysis is very data demanding, and not all proteins have many known analogs and homologs to carry out the CoC analysis. To this end, attempts are made to determine folding transition state(s) at the atomic level of resolution from molecular dynamics simulations of individual proteins.

MOLECULAR DYNAMICS: Folding Pathways Inferred from Unfolding Simulations

Molecular dynamics (MD) makes it possible to study proteins at very high space and time resolution: An all-atom protein model moves, according to Newtonian laws, in a liquid of explicit water molecules. However, a high price is paid for such resolution. Simulations of no longer than 10–50 ns can be performed at current computer power (with the exception of a single study that reached the 1000 ns landmark; 31). Although 10 ns in time may be sufficient to observe hydrophobic effects, electrostatic screening, and friction, it is clearly not long enough to simulate either folding or unfolding under natural or mildly denaturating conditions. To speed molecular events, extreme conditions of more than 500–1000 K and high pressure are applied. Clearly at such conditions (close to conditions in a gun shell during firing), only the unfolding of a protein can be observed. Then to infer folding from unfolding trajectories, one has to rely on microscopic reversibility and reverse sequence of events observed in unfolding.

In spite of these problems and the fundamental ambiguity of force fields, MD simulations of high temperature unfolding were able to recover coarse-grained sequence of folding events consistent with experimental data on CI2 (66, 68), SH3 (118), barnase (21), lysozyme (55), segment B1 from protein G (113), and other proteins (see 9, 20 for reviews). Unfortunately, most of these studies consider the sequence of (un)folding events on a scale of formation of secondary structure elements, melting of domains, or melting of the whole hydrophobic core, etc. Hence, contribution of an individual residue into kinetics of folding can hardly be evaluated directly.

However, Daggett and coworkers developed a method to reconstruct TSE from a small set of unfolding MD trajectories or even a single trajectory. The idea of the method is to find a region in the unfolding trajectory where the transition state is passed. Li & Daggett defined the transition state "as a small ensemble of structures populated immediately prior to the onset of a large structural change." To find the transition state, authors search for the "large structural change" in the unfolding trajectory: They project the multidimensional trajectory into two or three dimensions and then visually analyze the projection and protein structures along the trajectory. The "large structural change" is identified as a moment when

the trajectory goes from one cluster of points to the other in the projection and when the major weakening of the hydrophobic core is observed in the structure. As the authors state themselves this method "is neither precise nor rigorous, making extensive comparison to experiment imperative."

After the transition state is identified, Li & Daggett computed two quantities $\phi_M D$ and S that are compared with the experimental ϕ-values. Correlation of the structure index S with experimental ϕ-values for CI2 are reported to be 0.87. Note that the method of Li & Daggett relies heavily on the experimental information on the stage when the TSE is identified. So, it remains to be seen to what extent this method can predict experimental ϕ-values. The value of the MD, however, can be not in prediction of the experimentally observed quantities but in complementing the experimental information with the high accuracy structural details.

Li & Daggett analyzed a possible transition state for CI2 unfolding and suggested mutations that can stabilize the transition state and hence speed up folding. Ladurner et al tested these predictions with impressive results of substantial (up to 40 times!) increases in the rates of folding for the mutants (65). Importantly, most other mutations in CI2 slow down folding. To complete the cycle, the authors performed MD simulations of the mutants and identified their transition states.

Another MD study of CI2 was performed by Lazaridis & Karplus who simulated 24 unfolding trajectories. While emphasizing the diversity of unfolding trajectories, they found, in accord with experiment, the disruption of tertiary interactions between the helix and a two-stranded portion of the β-sheet as the primary unfolding event. It is important to note that these authors used experimental information about ϕ-values in CI2 in order to determine the transition state in their simulations. Thus, it is difficult to say whether the observed correlation with experiment is a mere consequence of this fact or if simulations provided some nontrivial results.

A different approach to the analysis of the TSE was developed by Pande & Rokhsar. They simulated unfolding of a small 16-residue β-hairpin peptide at a somewhat realistic temperature ($T = 400$) and identified several metastable states populated during unfolding. They mapped the TSE between these metastable states using p_{fold} analysis (see above). This approach is elegant, but it is not clear whether such a method can be used to determine TSE between globally folded and unfolded states in MD simulations, as no MD simulation is able to simulate full refolding. Another problem is that the metastable states that are relevant at low temperature folding events may not appear at high temperature unfolding trajectories (37a). The transition state itself can be sensitive to temperature, as it seems to be sensitive to denaturant concentration (36, 117). Despite these limitations, this new method of identifying transition state is a promising alternative to human- and experiment-guided approaches discussed above.

Success in reconstructing a folding picture consistent with experiment indicates that MD simulations can be helpful in interpreting experimentally observed ϕ-values and, perhaps, for their predictions in the future. The combination of experiment with MD simulation seems to be a promising approach, as MD complements

the experimental picture of the transition state with atomic details and short-time (on the scale of folding) dynamics.

One, however, needs to be very careful inferring the sequence of folding events or the structure of transition state from very few high temperature unfolding trajectories. As more studies are being done, another major limitation of MD becomes apparent. A limited number of MD trajectories (<25) precludes consistent comparison with experiment where all observations are averaged over a huge ensemble of folding proteins (37a). Recently Kazmirski et al studied several unfolding trajectories of BPTI, CI2, and barnase (56). These authors developed a variety of techniques to compare trajectories. The main conclusion of this study is that unfolding trajectories do not follow a narrow path from the native state, but are rather diverse. However, when structures sampled in unfolding trajectories were characterized by a small (<10) number of properties (for example, radius of gyration and helix content), a common path can be observed for some proteins. In BPTI, however, even the order of secondary structure melting was different in different MD simulations, requiring a decrease in the resolution of analysis to a few coarse-grained properties. The only scenario common to all three proteins was that unfolding started from expansion of the core followed by fraying of secondary structure elements leading to the transition state, in accord with earlier predictions from analytical theory (109). The transition state trajectories diverged very fast, sharing no commonality even in such a coarse-grained picture. This is consistent with the (inverse) description of the folding process that emerged in lattice simulation: stochastic search prior to the transition state (nucleus) and a directed pathway past the nucleus (2). Another lesson of the study (56) is the danger of projecting high-dimensional trajectories on a small number of dimensions (24, 25). While projections may appear similar, the actual trajectories can be very different and vice versa. The apparent similarity of trajectories can be a result of the selection of a projection axis that may not adequately describe the changes within the system.

In summary, while MD operates on a very high resolution protein model at a cost of necessity to simulate unfolding at extreme conditions. In order to have a consistent interpretation of these results from multiple trajectories, one needs to sacrifice high resolution and get back to a low-resolution protein needed. This problem suggests that intermediate-resolution models that allow one to study thousands of folding trajectories at normal conditions may be promising.

ALL-ATOM MONTE-CARLO SIMULATIONS: Possible Folding Mechanism at an Atomic Level of Detail

An intermediate resolution model has been developed recently by Shimada, Kussell, and Shakhnovich (unpublished data). In this approach, all heavy atoms are represented as interacting hard spheres of various sizes corresponding to their van der Waals radii. This model includes all degrees of freedom relevant to folding—all side chain and backbone torsions—and uses a Go atom-atom potential that makes

two atoms attract each other (when they are within a certain atom-type specific cutoff distance) if they are neighbors in the native state and repel each other otherwise. This energy function strongly favors the native state, making it the global energy minimum.

Folding dynamics is simulated using the Monte-Carlo technique. Using a small protein, crambin, 163 folding transitions from random coil to compact conformation, differing less than 1 Å rms. with the native state, were recorded. The disulfide bonds were treated as normal Go-interactions (at the beginning, the protein was fully unfolded). By recording many folding events over a wide range of temperatures, a possible folding mechanism for crambin is obtained. Folding occurs via a cooperative first-order process, and many folding pathways to the native state exist. One particular sequence of events constitutes a fast-folding pathway where kinetics traps (due to chain misfolding and sidechain mispacking) are avoided. This pathway includes formation of an α-helical hairpin followed by a rate-limiting step of nucleation of a β-sheet, propagating to the full native structure with concurrent side chain packing. These results present a proof-of-principle for the possibility of a solution of a protein folding problem at an all-atom level, provided that one has a realistic all-atom potential energy function that correctly favors the native state. Several other small proteins (SH3 domains and the IgB binding domain) have also been folded using this approach. These simulations may provide an insight into folding mechanisms of small proteins at the atomic level of resolution.

CONCLUDING REMARKS

The protein folding field has undergone an interesting evolution since its inception in the 1960s. First, biochemical thinking dominated, with each protein viewed as a unique system that required a full atomic-level description of its folding pathway. The folding pathway itself was viewed as a sequence of microscopically well-defined events akin to a simple chemical reaction.

This view changed in the early 1990s when simplified analytical and lattice models demonstrated their power to explain several key aspects of protein folding, such as cooperativity, nucleation, and resolving the earlier paradigm known as the "Levinthal Paradox." A "new view" [term suggested by R. Baldwin (119) in appreciation of lattice model simulations reported in (101)] of protein folding as a statistical process emerged. This makes folding akin to phase transition—an analogy explored by many researchers from both thermodynamic and kinetic perspectives (11, 38, 58, 70, 88, 93, 98, 104, 108, 109, 111). The phase transition analogy shifts the focus in addressing folding pathways from a microscopic, step-by-step description to the analysis of essential milestones on the pathway—the unfolded state, transition state, native state, and possible intermediates—viewing each of them as dynamic ensembles of conformations corresponding to local minima or saddle point(s) in the free-energy landscape. Those developments have brought experiment and theory closer together (35, 39). A key lesson from these studies

is the appreciation of certain aspects of universality in protein folding, suggesting that at a coarse-grained level of description, there is a small number of major scenarios and many proteins fall into one of them (79).

However, most recently, models were developed that made it possible to simulate the folding of small proteins at atomic or near-atomic levels of detail. The success of these simulations came partly from the enhanced power of computers and, to a great extent, from a better understanding of the general principles of protein folding. Probably, we are entering a new era of folding studies that will elevate our understanding of folding to the atomic level of detail and will finally result in a search strategy and energy function that are powerful enough to predict structure from sequence at an atomic resolution of 1–2 Å. This may finally render theoretical protein folding useful for application in functional genomics and drug design.

LITERATURE CITED

1. Abkevich V, Gutin A, Shakhnovich E. 1994. Free energy landscape for protein folding kinetics: intermediates, traps and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* 101:6052–62
2. Abkevich V, Gutin A, Shakhnovich E. 1994. Specific nucleus as the transition state for protein folding: evidence from the lattice model. *Biochemistry* 33:10026–36
3. Abkevich V, Gutin A, Shakhnovich E. 1995. Impact of local and non-local interactions on thermodynamics and kinetics of protein folding. *J. Mol. Biol.* 252:460–71
4. Alm E, Baker D. 1999. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. USA* 96:11305–10
5. Baker D. 2000. A surprising simplicity to protein folding. *Nature* 405:39–42
6. Baldwin R, Rose G. 1999. Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem. Sci.* 24:77–83
7. Bicout D, Szabo A. 2000. Entropic barriers, transition states, funnels, and exponential protein folding kinetics: a simple model. *Protein Sci.* 9:452–65
8. Boczko EM, Brooks CL. 1995. First-principle calculation of the folding free energy of a three-helix bundle protein. *Science* 269:393–96
9. Brooks CL. 1998. Simulations of protein folding and unfolding. *Curr. Opin. Struct. Biol.* 8:222–26
10. Bryngelson J, Wolynes P. 1987. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA* 84:7524–28
11. Bryngelson J, Wolynes P. 1990. A simple statistical field theory of heteropolymer collapse with application to protein folding. *Biopolymers* 30:177
12. Burton R, Myers J, Oas T. 1998. Protein folding dynamics: quantitative comparison between theory and experiment. *Biochemistry* 37:5337–43
13. Chan HS. 2000. Modelling protein density of states: additive hydrophobic effects are insufficient for calorimetric two-state cooperativity. *Proteins: Struct. Funct. Genet.* 40:543–71
14. Chan HS, Dill K. 1997. From levinthal to pathways to funnels. *Nat. Struct. Biol.* 4:10–19
15. Chiti F, Taddei N, White PM, Bucciantini

M, Magherini F, et al. 1999. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.* 6:1005–9

16. Choe S, Li L, Matsudaira P, Wagner G, Shakhnovich E. 2000. Differential stabilization of two hydrophobic cores in the transition state of the villin 14t folding reaction. *J. Mol. Biol.* 304:99–115
17. Clarke J, Cota E, Fowler S, Hamill S. 1999. Folding studies of immunoglobulin-like β-sandwich proteins suggest that they share a common folding pathway. *Fold. Des.* 7:1145–54
18. Clementi C, Nymeyer J, Onuchic J. 2000. Topological and energetic factors: What determines the structural details of the transition state ensemble and en-route intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298:937–53
19. Cregut D, Civera C, Macias M, Wallon G, Serrano L. 1999. A tale of two secondary structure elements: when a [beta] hairpin becomes an alpha-helix. *J. Mol. Biol.* 292:389–401
20. Daggett V. 1998. Long timescale simulations. *Curr. Opin. Struct. Biol.* 10:160–64
21. Daggett V, Li AJ, Fersht AR. 1998. Combined molecular dynamics and phi-value analysis of structure-reactivity relationships in the transition state and unfolding pathway of barnase: structural basis of hammond and anti-hammond effects. *J. Am. Chem. Soc.* 120:12740–54
22. Demirel MC, Atilgan AR, Jernigan RL, Erman B, Bahar I. 1998. Identification of kinetically hot residues in proteins. *Protein Sci.* 7:2522–32
23. Dill K. 1985. Theory for the folding and stability of globular proteins. *Biochemistry* 24:1501–19

23a. Dinner A, Karplus M. 2001. The roles of stability and contact order in determining protein folding rates. *Nat. Struct. Biol.* 8:21–22

24. Dinner A, Karplus M. 1999. Is protein unfolding the reverse of protein folding? A lattice simulation analysis. *J. Mol. Biol.* 292:403–19

24a. Dinner A, Karplus M. 1999. The thermodynamics and kinetics of protein folding: a lattice model analysis of multiple pathways with intermediates. *J. Phys. Chem.* 103:7976–94

25. Dinner A, Sali A, Smith L, Dobson C, Karplus M. 2000. Understanding protein folding via free-energy surfaces from theory and experiment. *Trends Biochem. Sci.* 25:331–39
26. Dodge C, Schneider R, Sander C. 1998. The hssp database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res.* 26:313–15
27. Dokholyan N, Buldyrev S, Stanley H, Shakhnovich E. 1998. Discrete molecular dynamics studies of the folding of a protein-like model. *Fold. Des.* 3:577–87
28. Dokholyan N, Buldyrev S, Stanley H, Shakhnovich E. 2000. Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol.* 296:1183–88
29. Deleted in proof
30. Du R, Pande V, Grosberg A, Tanaka T, Shakhnovich E. 1998. On the transition coordinate for protein folding. *J. Chem. Phys.* 108:334–50
31. Duan Y, Kollman P. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 282:740–43
32. Durr E, Jelesarov I, Bossard H. 1999. Extremely fast folding of a very stable leucine zipper with a strengthened hydrophobic core and lacking electrostatic interactions between helices. *Biochemistry* 38:870–80
33. Dyson HJ, Rance M, Houghten RA, Lerner RA, Wright PE. 1988. Folding of immunogenic peptide fragments of proteins in water solution. I. Sequence requirements for the formation of a reverse turn. *J. Mol. Biol.* 201:16–200

34. Dyson HJ, Rance M, Houghten RA, Wright PE, Lerner RA. 1988. Folding of immunogenic peptides-fragments of proteins in water solution. II. The rascent helix. *J. Mol. Biol.* 201:201–17
35. Fersht AR. 1995. Optimization of rates of protein folding: the nucleation-condenstation mechanism and its implications. *Proc. Natl. Acad. Sci. USA* 92:10869–73
36. Fersht AR. 1999. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*. San Francisco: Freeman
37. Fersht AR. 2000. Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl. Acad. Sci. USA* 97:1525–29
37a. Finkelstein AV. 1997. Can protein unfolding simulate protein folding? *Protein Eng.* 10:843–45
38. Finkelstein AV, Shakhnovich EI. 1989. Theory of cooperative transitions in protein molecules. II. Phase diagram for a protein molecule in solution. *Biopolymers* 28:1681–94
39. Galzitskaya OV, Finkelstein AI. 1999. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. USA* 96:11299–304
40. Deleted in proof
41. Grantchanova V, Riddle D, Santiago J, Baker D. 1998. Important role of hydrogen bonds in the structurally polarized transition state for folding of the src sh3 domain. *Nat. Struct. Biol.* 5:714–20
41a. Guerois R, Serrano L. 2000. The SH3-fold family: experimental evidence and prediction of variations in the folding pathways. *J. Mol. Biol.* 304:967–82
42. Gutin A, Abkevich V, Shakhnovich E. 1998. A protein engineering analysis of the transition state for protein folding: simulation in the lattice model. *Fold. Des.* 3:183–94
43. Deleted in proof
44. Hamill S, Steward A, Clarke J. 2000. The folding of an immunoglobulin-like greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* 297:165–68
45. Hao MH, Scheraga H. 1994. Monte-carlo simulation of a first order transition for protein folding. *J. Phys. Chem.* 98:4940–45
46. Hao MH, Scheraga H. 1994. Statistical thermodynamics of protein folding: sequence dependence. *J. Phys. Chem.* 98:9882–86
47. Hao MH, Scheraga H. 1998. Molecular mechanisms for cooperative folding of proteins. *J. Mol. Biol.* 277:973–83
48. Heinikoff S, Heinikoff J. 1993. Performance evaluation of aminoacid substitution matrices. *Proteins: Struct. Funct. Genet.* 17:49–61
49. Itzhaki L, Otzen D, Fersht A. 1995. The structure of the transition state for folding of chymotrypsin inhibitor 2 analyzed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* 254:260–88
50. Jackson S. 1998. How do small single-domain proteins fold? *Fold. Des.* 3:R81–91
51. Jackson S, Fersht A. 1991. Folding of chimotrypsin inhibitor 2. I. Evidence for a two-state transition. *Biochemistry* 30:10428–35
52. Karplus M, Shakhnovich E. 1992. In *Protein Folding*, 4:127–96. New York: Freeman
53. Karplus M, Weaver D. 1976. Protein-folding dynamics. *Nature* 160:404–6
54. Kaya H, Chan HS. 2000. Polymer principles of protein calorimetric two-state cooperativity. *Proteins: Struct. Funct. Genet.* 40:637–61
55. Kazmirski S, Daggett V. 1998. Non-native interactions in protein folding intermediates: molecular dynamics simulations of hen lysozyme. *J. Mol. Biol.* 284:793–806

56. Kazmirski S, Li A, Daggett V. 1999. Analysis methods for comparison of multiple molecular dynamics trajectories: applications to protein unfolding pathways and denatured ensembles. *J. Mol. Biol.* 290:283–304
57. Kim D, Gu H, Baker D. 1998. The sequences of small proteins are not extensively optimized for rapid folding by natural selection. *Proc. Natl. Acad. Sci. USA* 95:4982–86
57a. Kim DE, Yi Q, Gladwin ST, Goldberg JM, Baker D. 1998. The single helix in protein l is largely disrupted at the rate-limiting step in folding. *J. Mol. Biol.* 284:807–15
58. Klimov D, Thirumalai D. 1996. A criterion which determines foldability of proteins. *Phys. Rev. Lett.* 76:4070–73
59. Klimov D, Thirumalai D. 1998. Cooperativity in protein folding: from lattice models with sidechains to real proteins. *Fold. Des.* 3:127–39
60. Klimov D, Thirumalai D. 1998. Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism. *J. Mol. Biol.* 282:471–92
61. Koshi J, Goldstein R. 1997. Mutation matrices and physical-chemical properties: correlations and implications. *Proteins* 27:336–44
62. Kragelund B, Osmark P, Neergaard T, Schiodt J, Kristiansen K, et al. 1999. The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of acbp. *Nat. Struct. Biol.* 6:594–601
63. Kuhlman B, Baker D. 2000. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. USA* 97:10383–88
64. Ladunga I, Smith R. 1997. Amino acid substitutions preserve protein folding by conserving steric and hydrophobicity properties. *Protein Eng.* 10:187–96
65. Ladurner AG, Itzhaki LS, Daggett V, Fersht AR. 1998. Synergy between simulation and experiment in describing the energy landscape of protein folding. *Proc. Natl. Acad. Sci. USA* 95:8473–78
66. Lazaridis T, Karplus M. 1997. "New view" of protein folding reconciled with the old through multiple unfolding simulations. *Science* 278:1928–31
67. Leopold P, Montal M, Onuchic J. 1992. Protein folding funnels: a kinetic approach to the structure-sequence relationship. *Proc. Natl. Acad. Sci. USA* 89:8721–25
68. Li A, Daggett V. 1994. Characterization of the transition state of protein unfolding by use of molecular dynamics—chymostrypsin inhibitor-2. *Proc. Natl. Acad. Sci. USA* 91:10430–34
69. Li L, Mirny L, Shakhnovich E. 2000. Kinetics, thermodynamics and evolution of non-native interactions in protein folding nucleus. *Nat. Struct. Biol.* 7:336–41
70. Lifshitz IM, Grosberg A, Khohlov A. 1978. Some problems of statistical physics of polymers with volume interactions. *Rev. Mod. Phys.* 50:683–713
71. Lopez-Hernandez E, Serrano L. 1996. Structure of the transition state for folding of the 129 aa protein chey resembles that of a smaller protein, ci2. *Fold. Des.* 1:43–55
72. Lorch M, Mason JM, Clarke AR, Parker MJ. 1999. Effects of core mutations on the folding of a beta-sheet protein: implications for backbone organization in the i-state. *Biochemistry* 38:1377–85
73. Luisi D, Kuhlman B, Sideras K, Evans P, Raleigh D. 1999. Effects of varying the local propensity to form secondary structure on the stability and folding kinetics of a rapid folding mixed alpha/beta protein: characterization of a truncation mutant of the N-terminal domain of the ribosomal protein l9. *J. Mol. Biol.* 289:167–74
74. Main E, Fulton K, Jackson S. 1999. Folding pathway of fkbp12 and characterisation of the transition state. *J. Mol. Biol.* 291:429–44
75. Martinez J, Pissabarro T, Serrano L. 1998.

Obligatory steps in protein folding and the conformational diversity of the transition state. *Nat. Struct. Biol.* 5:721–29

75a. Martinez JC, Serrano L. 1999. The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nat. Struct. Biol.* 6:1010–16

76. Mateu M, del Pino MS, Fersht A. 1999. Mechanism of folding and assembly of a small tetrameric protein domain from tumour suppressor p53. *Nat. Struct. Biol.* 6:190–98

76a. McCallister EL, Alm E, Baker D. 2000. Critical roles of beta-hairpin formation in protein G folding. *Nat. Struct. Biol.* 7:669–73

77. Michnick S, Rosen MK, Wandless TJ, Karplus M, Schreiber SL. 1991. Solution structure of FKBP, a rotamase enzyme and receptor for FK506 and rapamycin. *Science* 252:836–39

78. Michnick S, Shakhnovich E. 1998. A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. *Fold. Des.* 3:239–51

79. Mirny L, Abkevich V, Shakhnovich E. 1996. Universality and diversity of the protein folding scenarios: a comprehensive analysis with the aid of lattice model. *Fold. Des.* 1:103–16

80. Mirny L, Abkevich V, Shakhnovich E. 1998. How evolution makes proteins fold quickly. *Proc. Natl. Acad. Sci. USA* 95:4976–81

81. Mirny L, Shakhnovich E. 1999. Universally conserved residues in protein folds. Reading evolutionary signals about protein function, stability and folding kinetics. *J. Mol. Biol.* 291:177–96

82. Moran L, Schneider J, Kentisis A, Reddy G, Sosnick T. 1999. Transition state heterogeneity in gcn4 coiled coil folding studied by using multisite mutations and crosslinking. *Proc. Natl. Acad. Sci. USA* 96:10699–704

83. Munoz V, Eaton W. 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA* 96:11311–16

84. Munoz V, Serrano L. 1996. Local versus nonlocal interactions in protein folding and stability—an experimentalist's point of view. *Fold. Des.* 1:R71–77

85. Myers J, Oas T. 1999. Reinterpretation of gcn4-p1 folding kinetics: partial helix formation precedes dimerization in coiled coil folding. *J. Mol. Biol.* 289:205–9

85a. Northrup SH, Pear MR, Lee CY, McCammon JA, Karplus M. 1982. Dynamical theory of activated processes in globular proteins. *Proc. Natl. Acad. Sci. USA* 79:4035–39

86. Oliveberg M, Tan Y, Silow M, Fersht A. 1998. The changing structure of the protein folding transition state: implications for the shape of the free-energy profile for folding. *J. Mol. Biol.* 277:933–43

87. Onuchic J, Socci N, Luthey-Schulten Z, Wolynes P. 1996. Protein folding funnels: the nature of the transition state ensemble. *Fold. Des.* 1:441–50

87a. Ortiz AR, Skolnick J. 2000. Sequence evolution and the mechanism of protein folding. *Biophys. J.* 79:1787–99

88. Pande VS, Grosberg AY, Rokshar D, Tanaka T. 1998. Pathways for protein folding: Is a "new view" needed? *Curr. Opin. Struct. Biol.* 8:68–79

89. Pande VS, Grosberg AY, Tanaka T. 1995. Freezing transition of random heteropolymers consisting of arbitrary sets of monomers. *Phys. Rev. E* 51:3381–92

89a. Pande VS, Rokhsar DS. 1999. Folding pathway of a lattice model for proteins. *Proc. Natl. Acad. Sci. USA* 96:1273–78

90. Pande VS, Rokhsar DS. 1999. Molecular dynamics simulations of unfolding and refolding of a beta-hairpin fragment of protein g. *Proc. Natl. Acad. Sci. USA* 96:9062–67

90a. Perl D, Welker C, Schindler T, Schroder

K, Marahiel, et al. 1998. Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nat. Struct. Biol.* 5:229–35

91. Plaxco K, Larson S, Ruczinski I, Riddle D, Buchwitz B, et al. 2000. Evolutionary conservation in protein folding kinetics. *J. Mol. Biol.* 298:303–12
92. Plaxco K, Simons K, Baker D. 1998. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* 277:985–94
93. Plotkin S, Onuchic J. 2000. Investigation of routes and funnels in protein folding by free energy functional methods. *Proc. Natl. Acad. Sci. USA* 97:6509–14
94. Portman J, Takada S, Wolynes P. 1997. Variational theory for site resolved protein folding free energy surfaces. *Phys. Rev. Lett.* 81:5237–40
95. Prieto J, Serrano L. 1997. C-capping and helix stability: the pro c-capping motif. *J. Mol. Biol.* 274:276–88
96. Privalov PL. 1989. Thermodynamic problems of protein structure. *Annu. Rev. Biophys. Biophys. Chem.* 18:1–47
97. Ptitsyn OB. 1998. Protein folding and protein evolution: common folding nucleus in different subfamilies of c-type cytochromes? *J. Mol. Biol.* 278:655–66
98. Ptitsyn OB, Kron A, Eizner YY. 1968. Simple statistical theory of self-organization of protein molecules. *J. Polym. Sci. C* 16:3509–16
99. Ptitsyn OB, Ting KLH. 1999. Non-functional conserved residues in globins and their possible role as a folding nucleus. *J. Mol. Biol.* 291:671–82
100. Ramanathan S, Shakhnovich E. 1994. Statistical mechanics of proteins with "evolutionary selected" sequences. *Phys. Rev. E* 50:1303–12

100a. Riddle DS, Grantcharova VP, Santiago JV, Alm E, Rucainski I, Baker D. 1999. Experiment and theory highlight role of native state topology in SH3 folding. *Nat. Struct. Biol.* 6:1016–24

101. Sali A, Shakhnovich E, Karplus M. 1994. How does a protein fold? *Nature* 369:248–51
102. Sfatos CD, Gutin AM, Shakhnovich EI. 1993. Phase diagram of random copolymers. *Phys. Rev. E* 48:465–75
103. Shakhnovich E. 1994. Proteins with selected sequences fold to their unique native conformation. *Phys. Rev. Lett.* 72:3907–10
104. Shakhnovich E. 1997. Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* 7:29–40
105. Shakhnovich E. 1998. Folding nucleus: specific or multiple? Insights from simulations and comparison with experiment. *Fold. Des.* 3:R108–11
106. Deleted in proof
107. Shakhnovich E, Abkevich V, Ptitsyn O. 1996. Conserved residues and the mechanism of protein folding. *Nature* 379:96–98
108. Shakhnovich E, Finkelstein A. 1982. On the theory of cooperative transitions in proteins. *Dokl. Acad. Nauk SSSR* 243:1247
109. Shakhnovich E, Finkelstein A. 1989. Theory of cooperative transitions in protein molecules. I. Why denaturation of globular protein is a first-order phase transition. *Biopolymers* 28:1667–81
110. Shakhnovich E, Gutin A. 1989. Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of replica approach. *J. Biophys. Chem.* 34:187–99
111. Shakhnovich E, Gutin A. 1993. Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci. USA* 90:7195–99
112. Shakhnovich E, Gutin A. 1993. A novel approach to design of stable proteins. *Protein Eng.* 6:793–800

113. Sheinerman FB, Brooks CL. 1998. Calculations on folding of segment b1 of streptococcal protein G. *J. Mol. Biol.* 278:439–56
114. Shoemaker B, Wang J, Wolynes P. 1999. Exploring structures in protein folding funnels with free energy functionals: the transition state ensemble. *J. Mol. Biol.* 287:675–94
115. Socci ND, Onuchic JN, Wolynes P. 1996. Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* 104:5860–68
116. Sosnick TR, Jackson S, Englander SW, DeGrado W. 1996. The role of helix formation in the folding of a fully α-helical coiled-coil. *Proteins* 24:427–33
116a. Taketomi H, Ueda Y, Go N. 1975. Studies on protein folding, unfolding and fluctuations by computer simulation. *Int. J. Pep. Protein. Res.* 7:445–49
117. Ternstrom T, Mayor U, Akke M, Oliveberg M. 1999. From snap-shot to movie: phi-value analysis of protein folding transition states taken one step further. *Proc. Natl. Acad. Sci. USA* 96:14854–59
118. Tsai J, Levitt M, Baker D. 1999. Hierarchy of structure loss in md simulations of src sh3 domain unfolding. *J. Mol. Biol.* 291:215–25
119. Udgaonkar J, Baldwin R. 1988. Nmr evidence for an early framework intermediate on the folding pathway of the ribonuclease a. *Nature* 335:694–700
120. Viguera A, Serrano L, Wilmanns M. 1996. Different folding transition states may result in the same native structure. *Nat. Struct. Biol.* 4:939–46
121. Villegas V, Martinez JC, Aviles FX, Serrano L. 1998. Structure of the transition state in the folding process of human procarboxypeptidase a2 activation domain. *J. Mol. Biol.* 283:1027–36
122. Zhou YQ, Karplus M. 1997. Folding thermodynamics of a model three-helix-bundle protein. *Proc. Natl. Acad. Sci. USA* 94:14429–32

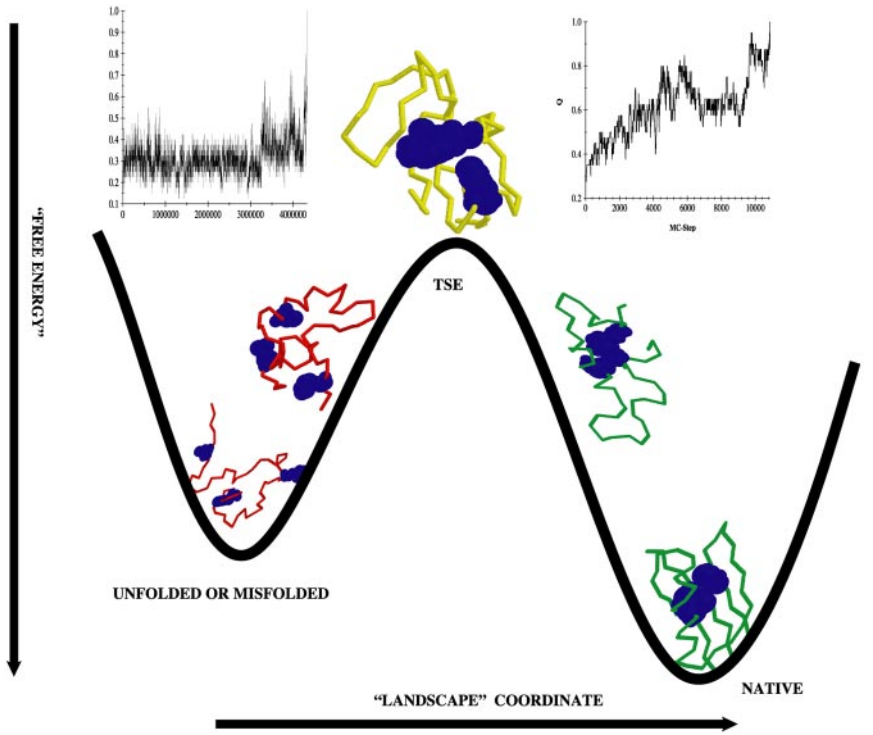

Figure 1 Specific nucleus mechanism of protein folding. The free energy landscape is shown schematically as one-dimensional. Nucleus residues are shown schematically as blue space-filling residues. Conformations that do not have the nucleus assembled are precritical (*shown in red*) and are committed to unfolding and long fluctuations in the unfolded state prior to the folding event (*upper left panel*). Conformations that are past the free energy barrier have the folding nucleus assembled. Those PCCs are committed to rapid continuous folding without further major free energy barriers (*right upper panel*). The separating set of conformations corresponding to the top of the free energy barrier is the transition state ensemble (TSE) whereby the nucleus is "almost formed". One of the conformations of the TSE is shown schematically in yellow. All representations in this figure are schematic. Each conformation shown (except the native state) is a member of a correspondingly broad ensemble. Inserts present degree of folding as a function of folding time for a typical folding trajectory.

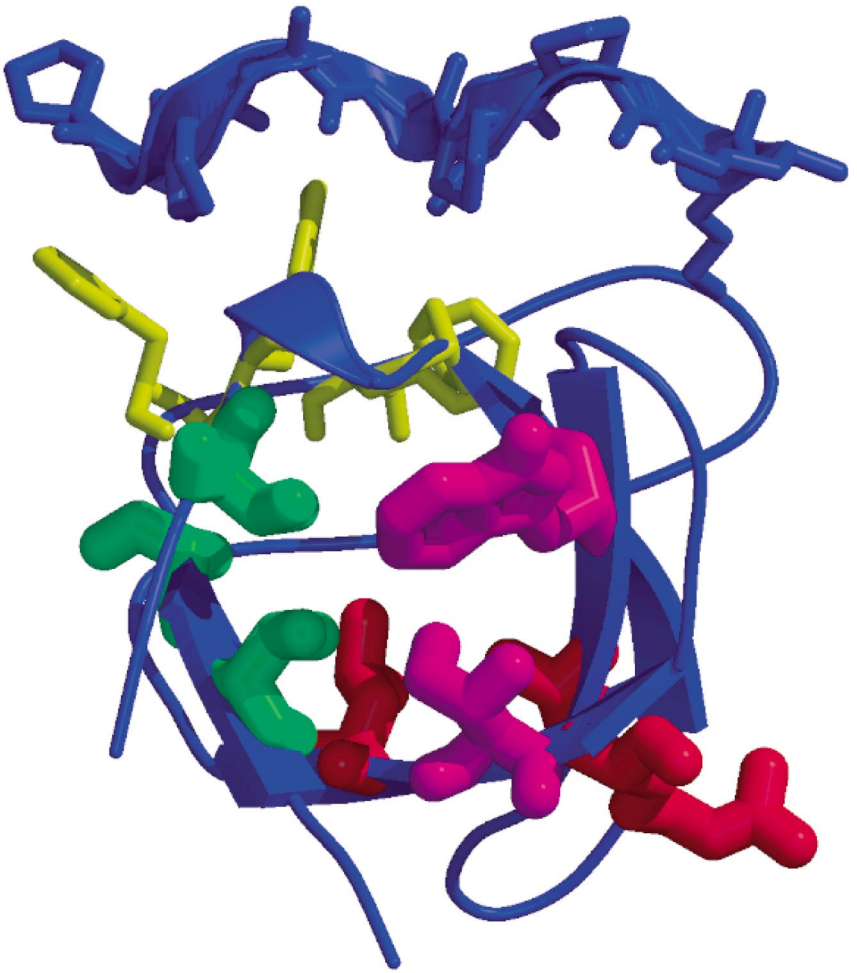

Figure 6 Structure of C-crk (*both in blue*) SH3 with proline peptide. Residues with high CoC and high conservatism across the families are shown in thick wire-frame. In magenta are residues with abnormal ϕ-values, I34 and W43. In red are residues L32, V35, and A45, all with high ϕ-values. In green are residues with low ϕ-values F10, A12, and V61. Residues that are not conserved between different families but have high CoC—Y14, F16, P57, and Y60—are shown in yellow. They contribute to the peptide binding pocket.