Understanding ensemble protein folding at atomic detail

IOP Publishing Journal of Physics: Condensed Matter

J. Phys.: Condens. Matter **20** (2008) 283101 (11pp) doi:10.1088/0953-8984/20/28/283101

TOPICAL REVIEW

Understanding ensemble protein folding at atomic detail

Stefan Wallin and Eugene I Shakhnovich[1]

Department of Chemistry and Chemical Biology, Harvard University, 12 Oxford Street,
Cambridge, MA 02138, USA

E-mail: eugene@belok.harvard.edu

Received 28 March 2008, in final form 4 April 2008
Published 2 June 2008
Online at stacks.iop.org/JPhysCM/20/283101

Abstract

Although far from routine, simulating the folding of specific short protein chains on the computer, at a detailed atomic level, is starting to become a reality. This remarkable progress, which has been made over the last decade or so, allows a fundamental aspect of the protein folding process to be addressed, namely its statistical nature. In order to make quantitative comparisons with experimental kinetic data a complete ensemble view of folding must be achieved, with key observables averaged over the large number of microscopically different folding trajectories available to a protein chain. Here we review recent advances in atomic-level protein folding simulations and the new insight provided by them into the protein folding process. An important element in understanding ensemble folding kinetics are methods for analyzing many separate folding trajectories, and we discuss techniques developed to condense the large amount of information contained in an ensemble of trajectories into a manageable picture of the folding process.

(Some figures in this article are in colour only in the electronic version)

Contents

1. Introduction

Akin to a phase transition rather than a classical chemical reaction, protein folding is the process that transforms a polypeptide chain from an unfolded, high-entropy state into its unique, native structure. The stability of the compact native state is maintained by many weak noncovalent interactions, dominated by hydrogen bonding and the hydrophobic effect. Hence, the formation of the native structure involves a competition between the loss of conformational entropy and the formation of favorable noncovalent interactions. This energy–entropy compensation which occurs during folding is large but remarkably balanced such that ultimately the native states of proteins are only marginally stable (typically 5–20 $k_{\mathrm{B}}T$ at room temperature where k_{B} is Boltzmann's constant and T the temperature in kelvin) [1, 2]. A large number of microscopically separate folding trajectories are possible in the transition from the unfolded to the native state, and understanding folding in atomic detail requires an ensemble approach.

[1] Author to whom any correspondence should be addressed.

Protein folding is of great biological and medical importance and a tremendous amount of effort has been put into understanding this process, both through theoretical approaches and *in vitro* experiments [3–7]. In the past, the focus on folding was primarily motivated by the observation that the biological function of a protein is determined not by its amino acid sequence directly but rather by the three-dimensional structure of the native state. Additional interest in folding has been spurred by two more recent observations, which in particular emphasize the importance of understanding folding as a *dynamic* process. First, it has become clear that a significant portion of proteins are partially, or even wholly, unfolded on their own and yet biologically active in some cases [8]. Folding may then occur as they bind to their target molecules, which establishes a direct link between the folding process and the biological function of the protein. Second, it has been found that 'errors' in the folding process, so-called misfolding, can have severe consequences for the host organism [9]. Parkinson's and type II diabetes are examples of diseases which have been linked to the misfolding and subsequent aggregation of protein chains. These types of protein conformational diseases are often particularly severe and debilitating. Understanding, in atomic detail, the folding and misfolding processes may be useful as a starting point for devising therapeutic strategies for this class of diseases [9].

Because of the potential for microscopic insight into the mechanism of folding, simulating protein folding on the computer at atomic detail has been a long-standing challenge in biology. Until very recently, performing such folding simulations was computationally prohibitively demanding. In the last few years, however, it has become feasible to obtain representative conformational sampling in atomic-level simulations for some small proteins and peptides, at least when these models are combined with minimalistic potential energy functions for describing the physical interactions that drive folding. This does not mean that the protein folding problem is solved, however. Rather, it means that the focus should be on developing potential energy functions that are simple yet physically sound, i.e. agree with available experimental data. The ability to make detailed, quantitative comparisons between simulation and experiment for specific proteins at the atomic level is an exciting recent development. Vital for the goal of 'calibrating' protein models to experimental thermodynamic and kinetic data are small autonomously folding proteins and peptides with different types of secondary and tertiary structure. Several such small folding units have been discovered in recent years. Examples include the B domain of protein A and the headpiece subdomain of the F-actin binding protein Villin, which both have been intensively studied by simulation and experiment. Moreover, re-engineered versions of small proteins, with one or several amino acid mutations, have been developed to achieve extremely high folding rates, even approaching the 'speed limit' for folding [10], in order to close to gap between experiment and the limits of simulations. This is particularly important for detailed molecular dynamics models with an explicit representation of the solvent atoms (water), which are computationally extremely demanding. However, due to large-scale distributed computing projects such as the 'folding@home' project [11], even such explicit-water simulations have recently reached single-trajectory timescales which are approaching, or even exceeding, the folding times of small and extremely fast-folding helical proteins [12]. In particular, this means that a comparison between molecular dynamics explicit-water models and simpler minimalistic all-atom models can be made in some cases.

The emerging ability to perform such large-scale folding simulations creates a need to develop tools to analyze and organize the data. To this end, various graph-theoretical tools and clustering procedures have been developed by several groups. Such tools can play important roles for folding simulation studies. Their main goal is to condense the information contained in many folding events into a coherent coarse-grained description of the folding process, but they have also been used in more direct ways, for example, by helping to identify particular states, such as folding intermediates and transition-state ensembles.

1.1. Why is explicit-chain simulation of protein folding necessary?

Before turning to the specific progress made in recent simulation studies of protein folding, it is useful to briefly consider the advantages and difficulties of studying the protein folding process using explicit-chain models. In order to trust the dynamical behavior of a chain model it is, of course, crucial that simulations are able to reproduce key aspects of the available experimental data for specific proteins. For some common types of experimental data this has proven to be challenging (although not impossible). It may in this situation be tempting to turn to other types of simple theoretical models which lack an explicit representation of the protein chain. Several such theoretical constructs [13–15], where the free-energy function is postulated and expressed in various ways using information from the native structure, have had limited success in reproducing some experimental data on folding, in particular the folding rate k_f. These results hinge mainly on an empirical observation that was made in 1998 by Plaxco *et al*, namely that for two-state proteins (most small single-domain proteins) the logarithm of the folding rate, $\ln k_\mathrm{f}$, correlates significantly with the so-called relative contact order, a simple topological parameter derived from the coordinates of the native structure which measures the average sequence separation between contacting residues relative to protein length [16]. However, as more data became available and all rather than only two-state proteins were included, other properties such as absolute contact order (which is closely correlated itself with protein length) showed much better predictive power while relative contact order exhibited an almost random correlation [17–20]. Despite the success of these simple theoretical constructs in reproducing certain experimental data, it has been argued that an explicit treatment of the protein chain is necessary to gain mechanistic insights into folding dynamics [21, 22]. Indeed, it can be difficult, or sometimes even misleading, to draw conclusions about the folding mechanics from the success of a simple model that does not take excluded-volume effects into account, as was discovered recently in a critical assessment of the 'topomer search model' for protein folding [23]. The attractiveness of an all-atom explicit-chain approach to protein folding is obvious, in that it can, at least in principle, provide an atomic-level description of the folding mechanics of a polypeptide chain from its unfolded state, through the transition state, and to the native structure. The disadvantage is that it is notoriously difficult to simulate folding on the computer. This difficulty notwithstanding, remarkable progress has been made the last 10 years or so, which we will discuss in some detail below.

2. Molecular dynamics physics-based protein models

The moderate stability of proteins are maintained by several different types of weak noncovalent interactions (one exception is the disulfide bond that can occur between pairs of cystein residues), such as van der Waals interactions, electrostatic interactions, desolvation effects, hydrogen bonding, and the hydrophobic effect. Because the origin of these different energetic effects are well understood in principle, a relatively straightforward approach to studying protein dynamics is to describe all (or most) contributing molecular forces based on a (classical) microscopic physical description with the model parameters obtained from detailed calculations (quantum mechanics). This is the general approach taken by popular standard force fields such as CHARMM and AMBER. Although simulations of protein dynamics using such explicit-water force fields can provide a measure of physical insight, their usefulness for protein folding is often limited by the computational demands and the uncertainty as to how accurately these empirical potentials can reproduce protein energetics. For these reasons, solvent molecules are often represented implicitly; for a review on recent advances in implicit solvent models, see [24]. It should be pointed out, however, that attempts at *ab initio* protein folding in explicit-water are being made by massive computational efforts. For example, Pande and co-workers recently obtained a large number of relatively short explicit solvent MD trajectories comprising a total of ≈500 μs simulation time for a double mutant of the helical protein Villin [12], which illustrates the current state-of-the-art in explicit-water simulation of protein folding. Despite the unprecedented computational effort, however, trajectories started from random coil structures very rarely reached the fully folded state. Recent explicit solvent simulations by Schulten and co-workers reached timescales of up to 10 ms—more than the folding time of the WW domain which they simulated [25]. The conformations reached were nevertheless dissimilar to the native conformation and the authors argued that the likely reason for that is in the inaccuracy of the CHARM22 force field used [25]. These results illustrate that computational resources are not the only hurdle to solving the protein folding problem and that even the detailed standard explicit-water force fields may need to be further developed in order to be widely applicable to folding studies.

3. Toward simple realistic all-atom models for protein folding

An alternative approach for capturing the different types of interactions governing protein dynamics (including solvent effects) in a combined way is by using the vast amount of information available in the many experimentally determined native structures of proteins. This is the central idea behind knowledge-based, or statistical, potentials. More precisely, the aim is to extract information about putative interaction energies between different residues from the pairing frequencies in known protein structures. Miyazawa and Jernigan [26] derived residue–residue energies using the quasichemical approximation which assumes that the probability of contact frequencies obey a Boltzmann distribution (the applicability of Boltzmann distribution to frequencies of contact pairs in proteins has been critically evaluated in [27]). The resulting so-called Miyazawa–Jernigan interaction potential matrix has been used in numerous applications, often together with simple lattice models for protein folding. The quasichemical approach has also been generalized to other types of more detailed interactions, including contact energies between individual atom groups [28], local chain propensities [29, 30], and hydrogen bonding [31, 32].

In our lab, we took a different approach and developed [33] an all-atom knowledge-based potential that does not rely on the quasichemical approximation. Instead, for two atom types A and B, the contact energy takes the form

$$E_{\mathrm{AB}} = \frac{-\mu N_{\mathrm{AB}} + (1-\mu)\tilde{N}_{\mathrm{AB}}}{\mu N_{\mathrm{AB}} + (1-\mu)\tilde{N}_{\mathrm{AB}}},$$

where $N_{\mathrm{AB}}\,(\tilde{N}_{\mathrm{AB}})$ is the number of instances where atom types A and B are found in contact (not found in contact) in the protein structure data set. The weighted average form of this 'μ-potential' was chosen such that E_{AB} coincides with the Go-potential when trained on a single protein structure and the number of atom types becomes the number of atoms in the protein. The μ-potential procedure ensures that all interaction energies E_{AB} lie between -1 and $+1$, which is an advantage over the quasichemical potential form which can overestimate the repulsion between atom types that are not observed to interact in the database. The parameter μ controls the overall tendency to obtain attractive or repulsive interaction energies. We have found that a suitable choice of μ is usually obtained by requiring that the average contact energy, taken over all-atom types, is zero, based on extensive test simulations performed on real protein sequences.

4. All-atom *ab initio* folding of diverse proteins

A prerequisite for studying the folding process at atomic detail is a model where dominant low-energy states resemble real native protein structures. To this end, we recently developed an all-atom model based on the μ-potential and applied it to a set of amino acid sequences corresponding to proteins with diverse secondary and tertiary structures [34]. The energy function U of this model is relatively simple with only three terms, $U = E_{\mathrm{con}} + aE_{\mathrm{trp}} + bE_{\mathrm{hb}}$, corresponding to a μ-potential contact term (E_{con}), a term that models sequence-based local conformational preferences (E_{trp}), and a term for hydrogen bonding which can be modulated using secondary structure prediction data (E_{hb}). Only three adjustable parameters exist in the model, a and b which determine the relative weights of the energy terms and a parameter β which controls the hydrogen-bonding strength in β-sheet conformations. These three parameters were adjusted based on the ability of the model to fold a training set of 10 proteins and were then held fixed for a test on a larger set of 13 proteins. Despite the simple form of the energy function, the energy-minimum structures obtained through replica-exchange Monte Carlo (REMC) simulations displayed strong similarities with the experimental structures, as is shown in figure 1. To ensure an effective conformational search for the minimum-energy state, the Monte Carlo procedure used in this study included a knowledge-based move set [35]. The main reason for including this move set was to enhance the sampling of relevant regions in the conformational space (e.g., α-helix and β-sheet regions) but it also meant that the results were not directly amenable to a physical interpretation of the folding trajectories. This sampling issue can be addressed by using a different choice of move set and, moreover, through the use of fixed temperature Monte Carlo simulations rather than REMC, the trajectories obtained could be used to study the actual folding dynamics. However, the goal here was mainly to study the structural characteristics of the low-energy states for a set of widely different sequences. An important implication of these results is that a knowledge-based transferable potential with a relatively simple form is sufficient to fold a diverse set of small proteins (based on the energy-minimum criterion) to at least moderate resolution, and in some cases high resolution. This also further emphasizes the importance of particular factors in the basic physics of folding, as well as for the prediction of the structure of small compact proteins, such as chain compaction resulting from hydrophobic residues, local sequence-dependent conformational preferences, and hydrogen bonding specific to α- and β-conformations.

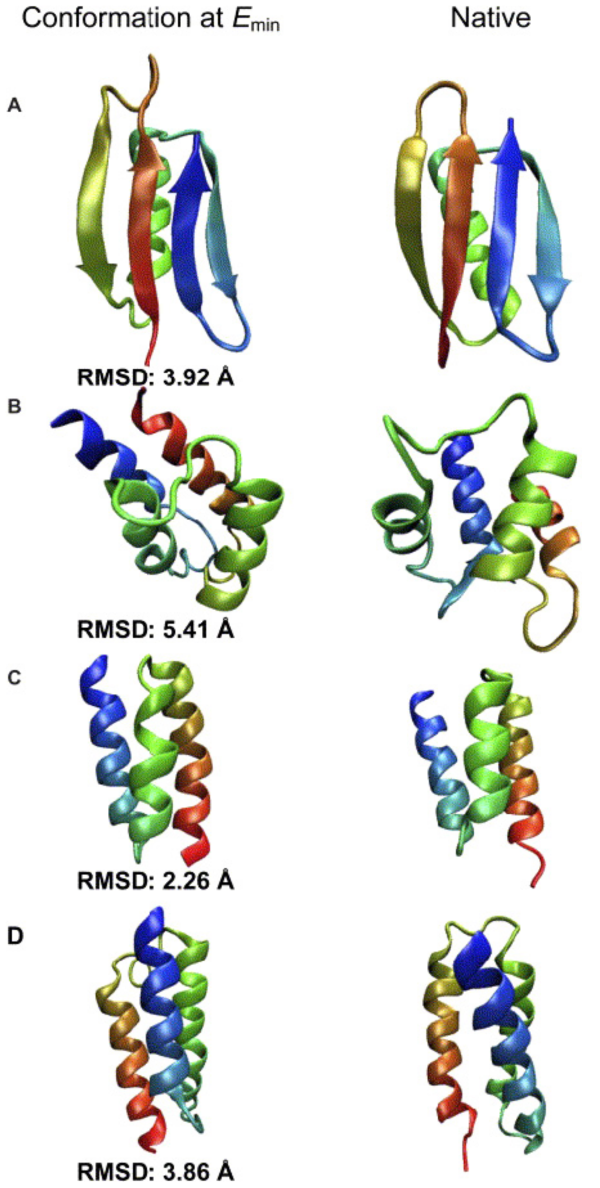

Figure 1. Comparison of the native structures of 4 proteins and the corresponding minimum energy (E_{min}) conformations obtained from REMC simulations of a simple transferable all-atom protein model [34]. RMSD values indicate the structural differences between native and E_{min} structures. The proteins are (A) IGG binding domain of protein G, (B) apo calbindin D9K, (C) albumin binding domain of protein G, and (D) a *de novo* designed protein model of a radical enzyme. Reprinted from [34]. Copyright 2007, with permission from Elsevier.

5. Detailed analysis of kinetic folding trajectories: clustering and structural graphs

With models for protein folding becoming more and more accurate in describing the native states of proteins as low-energy conformations, the ability to obtain large numbers of folding trajectories for particular proteins is also starting to become a reality. This situation then presents an interesting question: how can data from folding simulations be analyzed and organized into a coherent description of the folding process? It is true that not all aspects of folding kinetics require elaborate analysis methods. Relaxation behavior, chain collapse, formation of secondary structure, etc, can often be trivially obtained from an ensemble of folding trajectories. In fact, these relatively simple types of measurements can yield valuable insights into folding dynamics and are often useful for making direct comparisons with experimental data. However, from computer simulations of folding it is possible to obtain much more detailed kinetic information about the folding process, such as alternative folding 'pathways' or the structural characteristics of metastable states. After all, obtaining such kinetic information at an atomic-level resolution is one of the main goals of simulation studies. The use of graphs, network, and other clustering techniques to analyze folding trajectories has a relatively long history. The first to develop such a method was Levitt who introduced a pairwise distance matrix (the distances can for example be the C_{α} root-mean-square deviations (RMSDs) between the generated conformations) [36] and a procedure for projecting this high-dimensional matrix onto a two-dimensional plane such that the residual square error is minimized. Li and Daggett used this method for analyzing the unfolding of Chymotrypsin inhibitor 2 [37]. In particular, these authors attempted to locate the TSE by finding the first occurrence of the rapid structural changes which are expected to follow the passage across the transition state and into the unfolded ensemble. This turned out to be a challenging task as unfolding trajectories are usually recorded in MD simulations at very high temperatures and only gradual changes are discernible from such trajectories. Brooks and co-workers introduced a method for clustering conformations within trajectories based on a non-hierarchical clustering scheme and applied it to a short peptide [38]. Another approach to analyze large numbers of trajectories is to create a Markov state model (MSM) [39–42] which attempts to find transition probabilities (computed from the simulations) between states defined as conformational clusters. This approach holds the promise of being able to describe the transition between states through simple matrix multiplication operations, which makes it ideal for use in conjunction with large-scale distributed computing experiments which normally provide huge numbers of short trajectories. However, research into the accuracy of MSMs in describing the actual underlying dynamics is still ongoing [43].

In analyzing the simulated kinetics of a 20 amino acid three-stranded β-sheet peptide, beta3s, Rao and Caflisch developed a method for constructing a 'protein folding network' of kinetically linked conformations [44]. In this approach, an initial coarse-graining of the conformational space was performed such that each node in the graph represented a set of conformations with an identical secondary structure assignment. Undirected links were then placed between nodes based on the observation of actual transitions in single trajectories. The network resulting from their analysis of beta3s is presented in figure 2, where the native basin of attraction (FS) is shown at the bottom. Two parallel folding pathways are seen as two dense areas of transitions (links) with two different transition states, TSE1 and TSE2. By inspection of the transition-state structures, Rao and Caflisch could determine that the two pathways correspond to an initial formation of the C- and N-terminal hairpin of the three-stranded β-sheet peptide, respectively. Individual members of TSE1 and TSE2 were determined through p_{fold} analysis (p_{fold} analysis will be discussed in more detail below). Putative TS conformations were identified by selecting nodes with two properties which can be expected to characterize transition-state structures: (1) a high connectivity-to-statistical weight ratio and (2) a low clustering coefficient. An interesting finding by Rao and Caflisch is that calculated p_{fold}-values tended to be strongly correlated with the average neighbor connectivity, which suggests that an intelligent selection of putative TS structures can be made by considering properties of the network itself. Additional studies to elucidate this issue would be interesting because it might provide a means to speed up the identification of the TSE through p_{fold} analysis, which is computationally a quite demanding exercise.

A weakness in the protein network analysis of Rao and Caflisch is that the direction of time in the population of the various clusters cannot be clearly deduced. An approach that fully addresses this issue is the structural cluster analysis framework introduced recently by our group [45]. The aim this analysis is (1) to detect various trends among the many microscopic pathways taken in a large set of folding trajectories, and (2) to characterize these trends from both structural and kinetic perspectives. The fundamental idea behind the cluster analysis framework is the concept of a 'structural graph', which is schematically illustrated in figure 3. All conformations of the three fictitious trajectories in figure 3 are clustered together based on their pairwise structural similarities, creating various clusters with structurally coherent conformations. These clusters are informative by themselves as they demonstrate which structural motifs are common during folding but do not otherwise provide any kinetic information. We therefore introduced a quantity F, the flux of a cluster, which is defined as the fraction of all trajectories passing through the cluster, thus quantifying the cluster's kinetic significance. The native state, which can be identified as the giant component of the graph (GC, the largest cluster), will have $F = 1$ because all trajectories eventually reach N. In figure 3, this is the rightmost cluster. If other clusters with $F = 1$ are found, they can be interpreted as obligatory intermediate

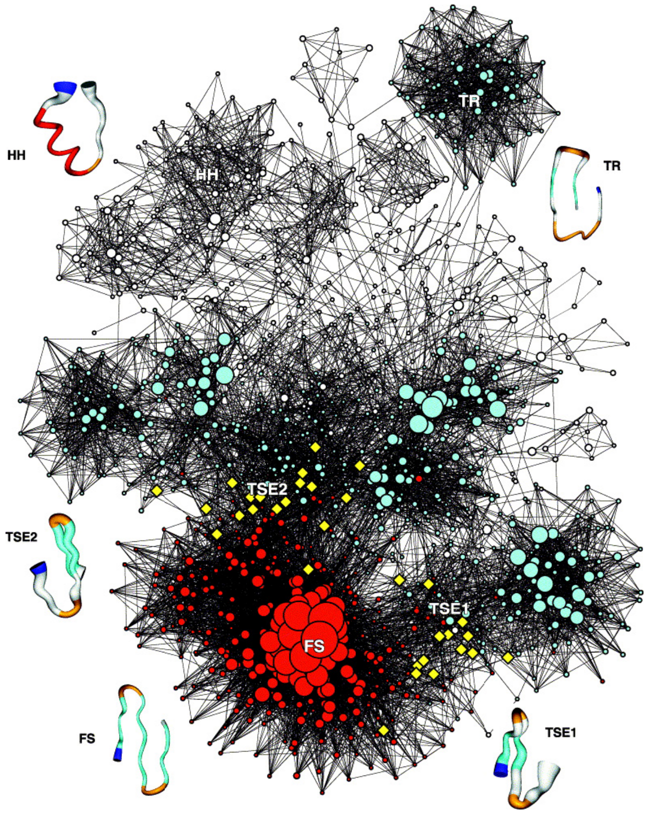

Figure 2. Protein folding network illustrating the conformational space and folding of the three-stranded peptide beta3s [44]. Nodes are clusters of identical secondary structure assignments and their sizes reflect their statistical weights (i.e. their free energies), and the node colors indicate average neighbor connectivities. Yellow diamonds are TS conformations identified through p_{fold} analysis. Two distinct transition-state regions, TSE1 and TSE2, emerge from the analysis. Reprinted from [44]. Copyright 2004, with permission from Elsevier.

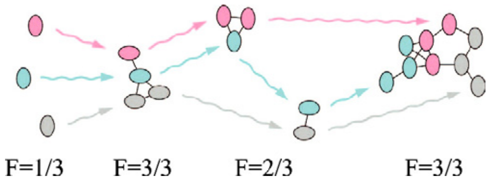

Figure 3. A schematic view explaining the concept of a structural graph and the flux, F. Each node (colored ovals) represents a single conformation and edges (solid lines) are drawn between structurally similar conformations (as determined by a criterion $d < d_c$, where d is a structural distance measure and d_c a selected cutoff value). Different colors indicate conformations from different trajectories and wavy lines indicate the direction of time, t. A collection of nodes that are linked by edges (either directly or in several steps) belong to the same cluster, and hence share strong structural similarities. Figure reprinted with permission from [45]. Copyright 2006 by the National Academy of Sciences.

states [45]. We have used this theoretical framework to analyze the folding process of several small helical proteins [45–47]. In particular, it was successfully used to study, in full atomic detail, the folding of the Engrailed Homeodomain, where several important states on the main pathway for folding, including collapsed partial helical states and an intermediate state, could be identified [45] (see figure 4). In this analysis, different structural similarity measures were used to create structural graphs, which was important as each measure could provide different structural insights into the folding process.

6. Three-helix-bundle folding kinetics: universality and diversity

As discussed above, one of the main goals of all-atom simulation of protein folding is to provide a description of the folding process at microscopic detail. This goal is closest to being met for small three-helix-bundle proteins [48], which are among the simplest protein folds that contain both secondary and tertiary structure. Folding simulations of this fold are simplified by the fact that α-helices are local structural elements and, as such, less complicated to model than β-sheets, which may involve chain segments that are globally separated in sequence. A few key questions about the folding of three-helix-bundle proteins are being addressed both through *in vitro* experiments and theory (of course, similar questions are being asked about protein folding in general). For example, what are the main 'pathways' for folding? This question entails determining the order in which key events occur during the folding process, such as chain collapse, helix formation, formation of intermediates, etc. Another important question is: what are the structural characteristics of the transition-state ensemble? Also, determining to what extent there is a universal mechanism by which folding occurs for three-helix-bundle proteins is of course a crucial question. Two small three-helix-bundle proteins which have played a particularly central role in addressing these different issues are the B domain of protein A and the Villin headpiece. Computational studies have been performed using various approaches with a range of modeling complexities, from simple C_α Go-type to all-atom explicit-water modeling, for both protein A [33, 46, 49–62] and Villin [56, 57, 63–67].

One specific detail, relating to the folding pathways of three-helix-bundle proteins, which has been discussed somewhat extensively in many investigations is the order of formation of individual α-helices during the folding of protein A. Bai *et al* [68] performed circular dichroism measurements on individual fragments of protein A and found that the C-terminal α-helix (helix 3) is the only one of the helices with some stability on its own. Consistent with this observation, many atomic-level simulations [53, 54, 56, 57, 60, 62] of protein A, as well as our recent study [47], have predicted that helix 3 is the most stable and the first to form during folding. This view was challenged, however, by a recent comprehensive ϕ-value analysis performed on protein A by Fersht *et al* [69, 70], which suggested that the TSE consists of a nearly fully formed helix 2, stabilized by hydrophobic interactions from helix 1, while helix 3 is mostly unstructured. As pointed out by Fersht *et al* [70], this picture of folding is consistent overall with the results obtained by an all-atom simulation study by Cheng *et al* [51]. Despite this, however, the transition-state ensemble obtained by Cheng *et al* did not lead to ϕ values that matched experimental values in a quantitative way. By contrast, such an agreement was achieved in our simulations of protein A [47] (see figure 5), despite the early formation of helix 3. Our simulation results suggested a folding scenario where helix 3 forms early in the folding process but otherwise interacts only weakly with the 'nucleus' of the transition state consisting of a relatively well-formed helical hairpin formed by helix 1 and 2. This proposed folding scenario is compelling partly because it may reconcile the apparent discrepancy between the ϕ-value analysis [69, 70] and circular dichroism measurements [68] on protein A.

The Villin headpiece is one of the fastest folding proteins discovered to date, which makes this protein an ideal candidate for testing the accuracy of molecular dynamics simulations.

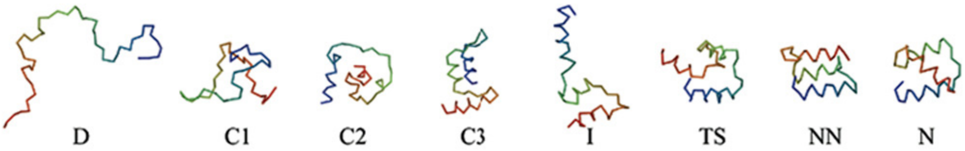

Figure 4. Folding from a denatured (D) state, which rapidly undergoes nonspecific collapse (C). There are several C states, characterized by increasing compaction and helical content. After the protein becomes sufficiently helical, the chain extends through fluctuations to an expanded intermediate (I) state, which allows rearrangement of the helices, and is followed by the transition state (TS). A final collapse to a near native (NN) state ensues, which proceeds through specific side chain packing and energetic relaxation to the native (N) state. C1, C2, C3, and I, may undergo rapid conversion.

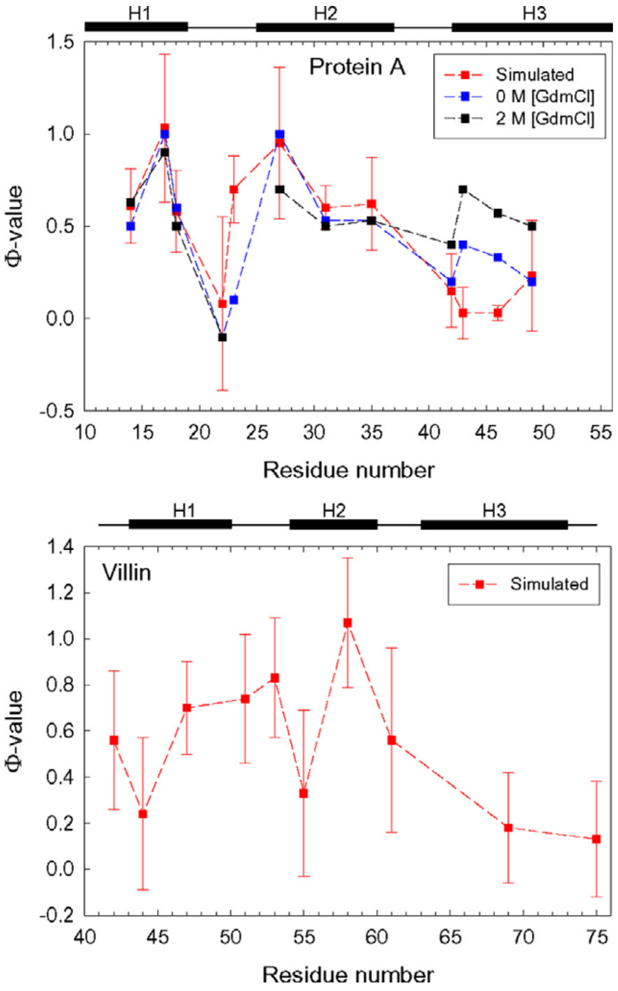

Figure 5. Comparison between simulated and experimental ϕ-values for the B domain of protein A and the ϕ-value prediction made for the Villin headpiece. Computational ϕ values are derived from a construction of the TSE of the two protein using a clustering procedure and p_{fold} analysis. Figure reprinted with permission from [47]. Copyright 2008 by the National Academy of Sciences.

Pande *et al* [67] used distributed computing to achieve thousands of trajectories of Villin with an implicit-water all-atom model. The length of each individual trajectory was very short (≈50 ns) compared to the overall folding time of Villin. These simulations are therefore mainly a characterization of the unfolded state under folding conditions and longer individual trajectories would be desirable. However, because of the large total simulation time (≈300 μs), a fraction of the trajectories are still expected to contain folding events which can then be further analyzed. In fact, this fraction allowed Pande *et al* to estimate of the folding rate of Villin to $5(+11, -3)$ μs. To experimentally test this prediction, Kubelka *et al* [71] used a laser-induced temperature jump to probe the kinetic behavior of a N27H mutant of Villin on the sub-microsecond timescale. The histidine mutation (H) was introduced to enhance the fluorescence signal of a solvent-exposed tryptophan residue which was used as a probe for folding. The folding time was measured to 4.3(±0.6) μs, in remarkably good agreement with the simulation results. This agreement does not guarantee, however, that the actual folding dynamics in the simulations is correct. In a further analysis of their trajectories, Pande *et al* observed an average rapid collapse of the chain with radius of gyration and native solvent accessibility close to that of the native state within 20 ns of simulation. In particular, this collapse meant that a natively solvent-exposed phenylalanine residue at position 36 (F36) was involved in a misfolded trapped state mediated by three other phenylalanines in the Villin sequence, which hampered the formation of the native structure. Hence, it was suggested that removing F36 would speed up folding. In response to this observation, Kubelka *et al* investigated the kinetics of the double mutant N27H/F36A but found no significant effect on the folding rate, suggesting that the misfolded trap observed in the simulations of Pande *et al* is not highly populated in reality or disassociates quickly enough to have little effect on the overall kinetics. One of the reasons for such discrepancies may be that folding events observed in very short individual trajectories represent anomalies not indicative of actual folding pathways. Paci and coauthors compared the folding of a relatively small beta-peptide observed in long trajectories with that of rare fast-folding events observed in short trajectories in distributed computing [72]. They showed that rare fast-folding events observed in distributed computing simulations are indeed atypical of a normal folding scenario [72]. Given the very small size of the proteins studied (in the range of 20–35 amino acid residues) and some arbitrariness in the definition of the folded state observed in the simulations a possibility exists that distributing computing generates many rapidly collapsed conformations and a few of them, by sheer chance, resemble the native state to a certain extent. A possible control for that would be a demonstration that the observed folding events are indeed sequence dependent (random collapse events apparently can occur with any sequence). However, to the best of our knowledge, such a control has not been carried out in distributed computing simulations.

In a recent effort to speed up the folding of Villin even further, Kubelka *et al* constructed another double mutant of Villin, where two buried lysine residues were substituted by norleucine residues, which was found to fold in the sub-microsecond folding range, making it an ultrafast folding protein [10]. This experimental advance, in combination with an unprecedented recent computational effort using distributed computing, allowed Pande *et al* to close the gap between *explicit-water* molecular dynamics simulation and experiment [12]. The authors obtained ≈500 independent simulation trajectories of each ≈1 μs initiated from 9 different conformations with varying degrees of residual native structure. Because of the extremely fast-folding nature of this Villin double mutant, each trajectory should be expected to contain on average at least one folding event in contrast to their earlier study [67]. Although foldings were observed for 3 of the initial conformations, which had the most structural similarity with the native conformations, very few of the simulations which started from >7 Å C_α RMSD away from the native structure resulted in folding. It should be pointed out though that a quite strict criterion was used to identify folded conformations. Nonetheless, these results clearly indicate that even current explicit-water force fields may require additional parameter tuning to accurately capture the complete folding dynamics of small proteins. Some folding trajectories were observed in [12], however, which were further analyzed. The relaxation behavior was found to be well described by a double-exponential function, both for an observable meant to mimic a tryptophan fluorescence signal and the native state population. Interestingly, this result is similar to what we observed for the Villin protein, suggesting that implicit solvent all-atom models may produce results similar to more elaborate explicit-water models.

Our recent study of protein A and Villin along with a previous investigation of the Engrailed Homeodomain [73] provide a comprehensive analysis of folding processes for three-helix-bundle proteins, within an implicit solvent approximation. In combining these results, we were able to formulate a universal picture of the folding of three-helix-bundle proteins [47]. This picture contains both universal features as well as significant diversity in the details. We find that the first step in folding is an initial collapse of the chain accompanied by partial formation of the α-helices (to a greater or lesser extent). On average, the chain thereafter remains relatively compact but frequent visits to more extended structures occur. During such fluctuations, the TSE may be located after which the chain collapses to the native state. The TSE consists of relatively well-formed helices organized into a two-helix hairpin and a third helix which is well formed but partially detached. Within this general framework there may be significant differences in the details, however. For example, the initial collapse phase can be accompanied by the formation of a single helix (such as helix 3 for protein A) or two helices (helix 2 and helix 3 for Villin). Moreover, there are two possibilities for the helical hairpin in the TSE, either involving helix 1 and 2 such as in protein A and Villin or helix 2 and 3 in Engrailed Homeodomain. Our analysis thus suggests that formation of a helix-turn-helix motif prior to entering the TSE might be a universal mechanism observed in folding of three-helix-bundle proteins, although details of which hairpin is formed may vary.

Many aspects of three-helix-bundle formation are clearly becoming known based on a combination of theoretical and experimental results. Still, some questions remain to be fully addressed. An issue that has only recently gained considerable attention is the character of the unfolded, or denatured, state ensemble. For a long time, the denatured state was generally thought of as an ensemble of unstructured, random coil conformations. This view is changing, however. With improvements in NMR spectroscopy and small-angle x-ray scattering in particular, the existence of secondary and sometimes even tertiary structure elements have been found in the denatured states of some proteins as well as in intrinsically disordered proteins [74]. The denatured states of protein A and Villin have so far been the focus of two computational studies [75, 76]. Another intriguing issue pertaining to the folding of three-helix-bundle proteins in general that has yet to be fully addressed is how nature selects the folding into one of the two possible mirror-image related three-helix-bundle topologies [77, 78], especially considering the very similar native contact pattern produced by the two topologies [79].

7. Folding thermodynamics: insights from peptide folding

In addition to kinetic data on proteins, equilibrium thermodynamic measurements on protein systems provide a way to test the physical accuracy of all-atom protein models. A complete thermodynamic characterization of protein folding is especially challenging to obtain computationally because it requires a representative sampling of the entire conformational space. Small peptide systems therefore provide a unique testing ground for comparing different protein models from a thermodynamic perspective, as well as an opportunity to gain insight into early events in the protein folding process [80]. In fact peptide folding includes many of the features observed in full-size protein folding, such as secondary structure formation, desolvation, ion pair formation, etc. Several peptides with around 20 amino acids or less have been discovered which fold into unique native states and are thus ideal test systems for protein models.

An illustrative example of the strength and limitations of current standard molecular dynamics simulations is given by the study of García and Sanbonmatsu [81] where the folding of the F_s peptide, a designed single-helix alanine/arginine 21 amino acid sequence, was studied with a replica-exchange molecular dynamics procedure and a modified version of the standard AMBER force field. Their simulations discovered an interesting detail in the peptide folding mechanism of the F_s peptide: the guanidinium group in the arginine sidechains may interact favorably with the carbonyl group four residues upstream in the chain and desolvate backbone hydrogen bonds, effectively increasing their strength and thus the overall stability of the α-helix. Despite such a detailed insight into the folding mechanism, a poor overall agreement with experimental thermodynamic data on the F_s peptide was obtained. A complete melting of the helix occurred only at temperatures $T \approx 500$ K. Similarly a weak T-dependence is in fact often observed in current molecular dynamics force fields and it has therefore been suggested that a reparametrization of these types of model might be needed [82]. It is interesting to compare these results with the thermodynamic behavior obtained by a minimalistic all-atom model developed by Irbäck and co-workers [83–85], with a force field based on effective hydrophobic forces and hydrogen bonding. Irbäck and Mohanty recently applied this model to a set of 5 peptides with diverse secondary structure contents, including the F_s peptide, and achieved accurate thermodynamic behavior, with melting temperatures in good *quantitative* agreement with experimental data, for all 5 peptides [83].

8. The transition-state ensemble in microscopic detail

A crucial part of achieving a complete ensemble view of any protein folding process is the characterization of the transition states that must be crossed during folding. Many small (and occasionally large [86]) single-domain proteins fold in an apparent two-state manner [87], i.e. with a single transition state. This means that folding proceeds more or less directly from the unfolded ensemble, U, to the native state, N, without significantly populating any intermediate state, at least not to an extent detectable by current experimental probes. A free-energy barrier separating the U and N states arises as the result of an imperfect cancelation between entropic and energetic contributions to folding, and the peak of this free-energy barrier is the transition state. The ensemble of conformations that make up the transition state is naturally of great importance in protein folding and a significant effort has been undertaken to try to characterize this state from different perspectives.

Because of its inherent unstable nature, the transition state is virtually impossible to study experimentally by direct means. However, ingenious indirect ways of probing the character of the transition state have been developed through protein engineering methods. One of these is ϕ-value analysis, pioneered by Fersht and co-workers [88], in which the effect of point mutations on the height of the free-energy barrier with respect to the unfolded and folded states, respectively, are probed (by measuring rate constants for folding and unfolding). The result of the analysis for an amino acid position i is encoded in a ϕ value, defined as $\phi_i = (\Delta\Delta G_{\mathrm{TS-U}}/\Delta\Delta G_{\mathrm{N-U}})$, where $\Delta\Delta G_{\mathrm{TS-U}}$ is the change in the free-energy difference between the TSE and U and $\Delta\Delta G_{\mathrm{N-U}}$ is the change in free-energy difference between the native state and U, resulting from the same mutation. A structural interpretation of ϕ values is straightforward for two ideal situations: $\phi_i = 0$ means that the residue i is as disordered in the TS as in U, and $\phi_i = 1$ means that i is as ordered in TS as in N. Intermediate ϕ values, which are normally obtained, are much more difficult to interpret, however. They indicate either that residue i is immersed in an environment that is partly native-like or that the obtained ϕ value is an ensemble average over multiple transition states resulting from parallel pathways for folding. Unconventional ϕ values (>1 or negative) sometimes also occur and are often taken as signs of nonnative interactions in the transition state. Moreover, the use of ϕ values derived from mutations with a small effect on protein stability ($\Delta\Delta G_{\mathrm{N-U}} < 1.7$ kcal mol^{-1}) is controversial [89, 90]. Despite these caveats ϕ-value analysis remains the most common experimental method for investigating the structural characteristics of the TSE for single-domain proteins.

Computer simulations present a unique opportunity to interpret ϕ values in structural ensemble terms at atomic resolution. Several groups have made attempts at combining protein engineering data and computer models to better understand the precise meaning of ϕ values and learn more about the TSE of particular proteins [91–94]. One approach aimed towards obtaining a detailed structural characterization of the TS was suggested by Vendruscolo and co-workers [93, 95], who introduced a pseudo-energy term involving a bias toward conformations which conform to experimentally determined ϕ values and applied their method to the 98-amino acid protein acylphosphatase. Conceptually, their procedure is similar to that used to generate native state structures compatible with NOE data from NMR experiments. Through these simulations, the authors were able to identify three residues (tyrosine 11, proline 54, and phenylalanine 94) in acylphosphatase which play key roles in organizing the polypeptide chain into its transition state—confirming the folding mechanism of nucleation via formation of a specific nucleus as was previously discovered in our lab [96, 97]. One potential weakness in the approach taken by Vendruscolo and co-workers is that computational ϕ values are interpreted in simple structural terms. The computational equivalent of ϕ values were calculated as the fraction of native contacts formed in the TSE relative to N, i.e. $\phi_i^{\mathrm{comp}} = \langle n_i^{\mathrm{TSE}}\rangle/n_i^{\mathrm{NAT}}$ where $\langle n_i^{\mathrm{TSE}}\rangle$ is the average number of native contacts in the TSE for a residue i and n_i^{NAT} is the number of native contacts. This definition has become the *de facto* standard in the computational protein folding literature but it is nonetheless an unverified assumption. In the absence of more exact ways to compute the free-energy contributions $\Delta\Delta G_{\mathrm{TS-U}}$ and $\Delta\Delta G_{\mathrm{N-U}}$ which make up the ϕ value quantity, ensembles obtained from ϕ value-restrained ensemble simulations should therefore be seen as *putative* TSEs rather than actual TSEs.

A way to construct a 'true' TSE is to verify whether generated conformations actually belong to the TSE. A rigorous such test is supplied by p_{fold} analysis which rests on the following simple observation about conformations C belonging to the TSE: independent trajectories passing through C will have an equal probability of first reaching the native state as the unfolded state, i.e. it will have a transmission coefficient, p_{fold}, of 0.5 [98]. This property can be viewed as a kinetic definition: conformations belonging to the TSE 'sit' at the top of the free-energy barrier (in fact at the saddle point region in the multidimensional conformational space separating the folded and unfolded basins of attraction). The quantity p_{fold} can be obtained by initiating a large number of folding simulations, with random initial conditions, from C and determining how many of the trajectories reach N without previously reaching U. Practically, however, a more convenient way to calculate p_{fold} is to determine the fraction of the trajectories that have 'committed' to folding after some time τ_{commit} [99]. An important question here is, of course, if the p_{fold} criterion should be imposed on a carefully constructed putative TSE? In studying the structural details of TSE for protein G using an all-atom Go-type model, Hubner *et al* [99] first generated ϕ value-restrained conformations, following the method of Vendruscolo *et al* [95], and then further tested the obtained structures through p_{fold} analysis. The study yielded some important observations. It was found that a gradual addition of ϕ values as restraints meant that the average p_{fold} value over all conformations in the putative TSE first grows and then saturates at ≈0.5. The putative TSE is therefore a reasonable approximation to the true TSE *on average*. However, the distribution of individual p_{fold} values was found to be distinctly bimodal, with most values close to either 0 or 1. This result is perhaps not surprising given that the transition state is an inherently unstable state and even small 'perturbations' away from the free-energy barrier may easily lead to conformations that are committed either to folding or unfolding. The bimodal distribution of p_{fold} values clearly indicates that no simple structural proxy exists that can detect true TS conformations with high accuracy.

However, some limitations to p_{fold} analysis method exist. Determining the p_{fold} for a single conformation requires many independent auxiliary simulations, making $p_{\rm fold}$ calculations computationally very demanding. It has also been pointed out that because force fields are not perfect, the estimation of $p_{\rm fold}$ values are susceptible to errors. Less-than-perfect correlations between $p_{\rm fold}$ values were indeed found when calculated using a range of different force fields [100]. Hence, there are significant technical challenges with $p_{\rm fold}$ analysis. However, it is also clear that unverified putative TSEs can sometime be misleading about the characteristics of the transition state. An illustration of this is provided by the SH3 domain, where a highly native-like topology of the TSE was obtained from (unverified) ϕ values constraints simulations [101, 102], while further $p_{\rm fold}$ analyses of the same protein domain have revealed a much richer picture of the TSE, with a highly polarized small nucleus but with significant parts of the chain left unstructured [91, 103].

9. Conclusion

Advances made over the last decade in protein modeling, along with increased computational resources and technologies, have made the folding of protein chains on the computer possible in certain cases. This has allowed an ensemble picture of the folding process to be constructed for small proteins, in particular three-helix-bundle proteins, through the generation of large numbers of complete folding trajectories at atomic detail. It is encouraging to see that many of these advances have been achieved with relatively simple models which combine relatively simple statistical potentials and all-atom representation of the protein chain. This further emphasizes the role of a few key physical aspects of the forces that drive folding, such as hydrogen bonding and the hydrophobic effect. Still, it is clear that further development and refinement of the existing models is needed, as well as comparative studies of different modeling approaches. In working toward elucidating the mechanics of folding, the importance of combining experimental and theoretical approaches cannot be overstated. Experimental data play key roles in verifying protein models, but true synergy between theory and experiments is also being obtained. For example, protein engineering experiments are combined with explicit-chain simulations thus achieving atomistically detailed constructions of transition-state ensembles for folding. Computer simulation folding, with its ability to provide microscopic insights, is being used to explain or even reconcile kinetic experimental data on folding which may, at first, appear contradictory. It is widely believed that an eventual solution to the protein folding problem will take a combined effort between simulation and experiment. Given the rapid development in both areas, fundamental additional progress in our understanding of the kinetics and thermodynamics of protein folding is bound to follow in the near future.

References

[1] Zeldovich K B, Chen P and Shakhnovich E I 2007 *Proc. Natl Acad. Sci. USA* **104** 16152

[2] Kumar M D, Bava K A, Gromiha M M, Prabakaran P, Kitajima K, Uedaira H and Sarai A 2006 *Nucleic Acids Res.* **34** D204

[3] Shakhnovich E I 2006 *Chem. Rev.* **106** 1559

[4] Eaton W A, Munoz V, Hagen S J, Jas G S, Lapidus L J, Henry E R and Hofrichter J 2000 *Annu. Rev. Biophys. Biomol. Struct.* **29** 327

[5] Thirumalai D and Hyeon C 2005 *Biochemistry* **44** 4957

[6] Pande V S, Grosberg A, Tanaka T and Rokhsar D S 1998 *Curr. Opin. Struct. Biol.* **8** 68

[7] Fersht A 1999 *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (New York: Freeman)

[8] Dyson H J and Wright P E 2002 *Curr. Opin. Struct. Biol.* **12** 54

[9] Chiti F and Dobson C M 2006 *Annu. Rev. Biochem.* **75** 333

[10] Kubelka J, Chiu T K, Davies D R, Eaton W A and Hofrichter J 2006 *J. Mol. Biol.* **359** 546

[11] Pande V S, Baker I, Chapman J, Elmer S P, Khaliq S, Larson S M, Rhee Y M, Shirts M R, Snow C D, Sorin E J and Zagrovic B 2003 *Biopolymers* **68** 91

[12] Ensign D L, Kasson P M and Pande V S 2007 *J. Mol. Biol.* **374** 806

[13] Alm E and Baker D 1999 *Proc. Natl Acad. Sci. USA* **96** 11305

[14] Garbuzynskiy S O, Finkelstein A V and Galzitskaya O V 2005 *Mol. Biol.* **39** 906

[15] Munoz V and Baker D 1999 *Proc. Natl Acad. Sci. USA* **96** 11311

[16] Plaxco K W, Simons K T and Baker D 1998 *J. Mol. Biol.* **277** 985

[17] Ivankov D N, Garbuzynskiy S O, Alm E, Plaxco K W, Baker D and Finkelstein A V 2003 *Protein Sci.* **12** 2057

[18] Kuznetsov I B and Rackovsky S 2004 *Proteins* **54** 333

[19] Mirny L and Shakhnovich E 2001 *Annu. Rev. Biophys. Biomol. Struct.* **30** 361

[20] Gromiha M M and Selvaraj S 2001 *J. Mol. Biol.* **310** 27

[21] Karanicolas J and Brooks C L 2003 *Proteins: Struct. Funct. Genet.* **53** 740

[22] Kaya H and Chan H S 2000 *Proteins: Struct. Funct. Genet.* **40** 637

[23] Wallin S and Chan H S 2005 *Protein Sci.* **14** 1643

[24] Feig M and Brooks C L 2004 *Curr. Opin. Struct. Biol.* **14** 217

[25] Freddolino P L, Liu F, Gruebele M H and Schulten K 2008 *Biophys J.*

[26] Miyazawa S and Jernigan R L 1996 *J. Mol. Biol.* **256** 623

[27] Finkelstein A V, Gutin A M and Badretdinov A 1995 *Subcell Biochem.* **24** 1

[28] Lu H and Skolnick J 2001 *Proteins: Struct. Funct. Genet.* **44** 223

[29] Godzik A and Skolnick J 1992 *Proc. Natl Acad. Sci. USA* **89** 12098

[30] DeWitte R S and Shakhnovich E I 1994 *Protein Sci.* **3** 1570

[31] Nishikawa K and Matsuo Y 1993 *Protein Eng.* **6** 811

[32] Kortemme T, Morozov A V and Baker D 2003 *J. Mol. Biol.* **326** 1239

[33] Kussell E, Shimada J and Shakhnovich E I 2002 *Proc. Natl Acad. Sci. USA* **99** 5343

[34] Yang J S, Chen W W, Skolnick J and Shakhnovich E I 2007 *Structure* **15** 53

[35] Chen W W, Yang J S and Shakhnovich E I 2006 *Proteins: Struct. Funct. Genet.* **66** 682

[36] Levitt M 1983 *J. Mol. Biol.* **168** 621

[37] Li A and Daggett V 1994 *Proc. Natl Acad. Sci. USA* **91** 10430

[38] Karpen M E, Tobias D J and Brooks C L 1993 *Biochemistry* **32** 412

[39] Andrec M, Felts A K, Gallicchio E and Levy R M 2005 *Proc. Natl Acad. Sci. USA* **102** 6801

[40] Jayachandran G, Vishal V and Pande V S 2006 *J. Chem. Phys.* **124** 164902

[41] Singhal N and Pande V S 2007 *J. Chem. Phys.* **126** 244101

[42] Swope W C, Pitera J W, Suits F, Pitman M, Eleftheriou M, Fitch B G, Germain R S, Rayshubski A, Ward T J C, Zhestkov Y and Zhou R 2004 *J. Phys. Chem.* B **108** 6582

[43] Swope W C, Pitera J W and Suits F 2004 *J. Phys. Chem.* B **108** 6571

[44] Rao F and Caflisch A 2004 *J. Mol. Biol.* **342** 299

[45] Hubner I A, Deeds E J and Shakhnovich E I 2006 *Proc. Natl Acad. Sci. USA* **103** 17747

[46] Hubner I A, Deeds E J and Shakhnovich E I 2005 *Proc. Natl Acad. Sci. USA* **102** 18914

[47] Yang J S, Wallin S and Shakhnovich E I 2008 *Proc. Natl Acad. Sci. USA* **105** 895

[48] Wolynes P G 2004 *Proc. Natl Acad. Sci. USA* **101** 6837

[49] Berriz G F and Shakhnovich E I 2001 *J. Mol. Biol.* **310** 673

[50] Boczko E M and Brooks C L III 1995 *Science* **269** 393

[51] Cheng S, Yang Y, Wang W and Liu H 2005 *J. Phys. Chem.* B **109** 23645

[52] Favrin G, Irbäck A and Wallin S 2002 *Proteins: Struct. Funct. Genet.* **47** 99

[53] García A E and Onuchic J N 2003 *Proc. Natl Acad. Sci. USA* **100** 13898

[54] Ghosh A, Elber R and Scheraga H A 2002 *Proc. Natl Acad. Sci. USA* **99** 10394

[55] Guo Z, Brooks C L III and Boczko E M 1997 *Proc. Natl Acad. Sci. USA* **94** 10161

[56] Jang S, Kin E, Shin S and Pak Y 2003 *J. Am. Chem. Soc.* **125** 14841

[57] Kim S-Y, Lee J and Lee J 2004 *J. Chem. Phys.* **120** 8271

[58] Linhananta A and Zhou Y 2002 *J. Chem. Phys.* **117** 8983

[59] Liwo A, Khalili M and Scheraga H A 2005 *Proc. Natl Acad. Sci. USA* **102** 2362

[60] St-Pierre J-F, Mousseau N and Derreumaux P 2008 *J. Chem. Phys.* **128** 045101

[61] Zhou Y and Karplus M 1999 *Nature* **401** 400

[62] Alonso D O V and Daggett V 2000 *Proc. Natl Acad. Sci. USA* **97** 133

[63] De Mori G M, Colombo G and Micheletti C 2005 *Proteins* **58** 459

[64] Duan Y and Kollman P A 1998 *Science* **282** 740

[65] Fernandez A, Shen M Y, Colubri A, Sosnick T R, Berry R S and Freed K F 2003 *Biochemistry* **42** 664

[66] Herges T and Wenzel W 2005 *Structure* **13** 661

[67] Zagrovic B, Snow C D, Shirts M R and Pande V S 2002 *J. Mol. Biol.* **323** 927

[68] Bai Y, Karimi A, Dyson J and Wright P E 1997 *Protein Sci.* **6** 1449

[69] Sato S, Religa T L, Daggett V and Fersht A R 2004 *Proc. Natl Acad. Sci. USA* **101** 6952

[70] Sato S, Religa T L and Fersht A R 2006 *J. Mol. Biol.* **360** 850

[71] Kubelka J, Eaton W A and Hofrichter J 2003 *J. Mol. Biol.* **329** 625

[72] Paci E, Cavalli A, Vendruscolo M and Caflisch A 2003 *Proc. Natl Acad. Sci. USA* **100** 8217

[73] Hubner I A, Deeds E J and Shakhnovich E I 2006 *Proc. Natl Acad. Sci. USA* **103** 17747

[74] Mittag T and Forman-Kay J D 2007 *Curr. Opin. Struct. Biol.* **17** 3

[75] Chowdhury S, Lei H and Duan Y 2005 *J. Phys. Chem.* B **109** 9073

[76] Jayachandran G, Vishal V, García A E and Pande V S 2007 *J. Struct. Biol.* **157** 491

[77] Irbäck A, Sjunnesson F and Wallin S 2000 *Proc. Natl Acad. Sci. USA* **97** 13614

[78] Regan L and Degrado W F 1988 *Science* **241** 976

[79] Wallin S, Farwer J and Bastolla U 2002 *Proteins: Struct. Funct. Genet.* **50** 144

[80] Gnanakaran S, Nymeyer H, Portman J, Sanbonmatsu K Y and García A E 2003 *Curr. Opin. Struct. Biol.* **13** 168

[81] García A E and Sanbonmatsu K Y 2002 *Proc. Natl Acad. Sci. USA* **99** 2782

[82] Zhou R, Berne B J and Germain R S 2001 *Proc. Natl Acad. Sci. USA* **98** 14931

[83] Mohanty S and Irbäck A 2005 *Biophys. J.* **88** 1560

[84] Irbäck A and Mohanty S 2006 *J. Comput. Chem.* **27** 1548

[85] Irbäck A, Samuelsson B, Sjunnesson F and Wallin S 2003 *Biophys. J.* **85** 1466

[86] Jones K and Wittung-Stafshede P 2003 *J. Am. Chem. Soc.* **125** 9606

[87] Jackson S E 1998 *Fold. Des.* **3** R81

[88] Ladurner A G, Itzhaki L S and Fersht A R 1997 *Fold. Des.* **2** 363

[89] Fersht A R and Sato S 2004 *Proc. Natl Acad. Sci. USA* **101** 7976

[90] Sanchez I E and Keifhaber T 2003 *J. Mol. Biol.* **334** 1077

[91] Ding F, Guo W, Dokholyan N V, Shakhnovich E I and Shea J-E 2005 *J. Mol. Biol.* **350** 1035

[92] Ladurner A G, Itzhaki L S, Daggett V and Fersht A R 1998 *Proc. Natl Acad. Sci. USA* **95** 8473

[93] Paci E, Vendruscolo M, Dobson C M and Karplus M 2002 *J. Mol. Biol.* **324** 151

[94] Settanni G, Gsponer J and Caflisch A 2004 *Biophys. J.* **86** 1691

[95] Vendruscolo M, Paci E, Dobson C M and Karplus M 2001 *Nature* **409** 641

[96] Abkevich V I, Gutin A M and Shakhnovich E I 1994 *Biochemistry* **33** 10026

[97] Li L and Shakhnovich E I 2001 *Proc. Natl Acad. Sci. USA* **98** 13014

[98] Du R, Pande V S, Grosberg A Y, Tanaka T and Shakhnovich E I 1998 *J. Chem. Phys.* **108** 334

[99] Hubner I A, Shimada J and Shakhnovich E I 2004 *J. Mol. Biol.* **336** 745

[100] Rhee Y M and Pande V S 2006 *Chem. Phys.* **323** 66

[101] Wright C F, Lindorff-Larsen K, Randles L G and Clarke J 2003 *Nat. Struct. Biol.* **10** 658

[102] Lindorff-Larsen K, Vendruscolo M, Paci E and Dobson C M 2004 *Nat. Struct. Mol. Biol.* **11** 443

[103] Hubner I A, Edmonds K A and Shakhnovich E I 2005 *J. Mol. Biol.* **349** 424