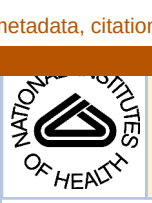



Published in final edited form as:
*Arch Biochem Biophys*. 2008 January 1; 469(1): 4–19.

# Protein Folding: Then and Now

**Yiwen Chen**[*], **Feng Ding**[*], **Huifen Nie**[*], **Adrian W. Serohijos**[*], **Shantanu Sharma**[*], **Kyle C. Wilcox**[*], **Shuangye Yin**[*], and **Nikolay V. Dokholyan**[*,†]
*Department of Biochemistry and Biophysics, The University of North Carolina at Chapel Hill, School of Medicine, Chapel Hill, NC 27599*

## Abstract

Over the past three decades the protein folding field has undergone monumental changes. Originally a purely academic question, how a protein folds has now become vital in understanding diseases and our abilities to rationally manipulate cellular life by engineering protein folding pathways. We review and contrast past and recent developments in the protein folding field. Specifically, we discuss the progress in our understanding of protein folding thermodynamics and kinetics, the properties of evasive intermediates, and unfolded states. We also discuss how some abnormalities in protein folding lead to protein aggregation and human diseases.

## I. Introduction

Protein folding refers to the process by which a protein assumes its characteristic structure, known as the native state. The most fundamental question of how an amino acid sequence specifies both a native structure and the pathway to attain that state has defined the protein folding field. Over more than four decades the protein folding field has evolved (Fig. 1), as have the questions pertaining to it. This evolution can be divided into two predominant phases. During the first phase, research was focused on understanding the mechanisms of protein folding and uncovering the fundamental principles that govern the folding transition. While the first phase provided general answers to the protein folding question, new and no less ambitious questions arose: what are the mechanisms of protein folding in a context, such as under the influence of other biological molecules in the cellular environment? This next set of questions defined the second phase in protein folding field evolution.

The first phase is akin to a romantic stage of research, where the final goal of studies may not be directly applicable to a broader understanding, or exploitable in a relevant science. The final goal is to determine the basic principles that relate protein sequence and structure. The second phase is a more pragmatic stage of research, where the applications drive research in the field and the rational manipulation of derived knowledge allows engineering of tools for advancement of a relevant science. For example, understanding the functional intermediates that accompany the transition of a protein *en route* to its native state may allow rational manipulation of protein structure via protein design. This example not only relates protein sequence, structure and function, but also demonstrates the engineering aspect of the modern protein folding field.

† To whom correspondence should be addressed: Nikolay V. Dokholyan, Department of Biochemistry and Biophysics, University of North Carolina at Chapel Hill, North Carolina 27599. Fax: 919-966-2852. Email: dokh@med.unc.edu.
[*]All authors contributed equally.

Next, we survey the questions of the modern protein folding field. We attempt to describe a number of directions where understanding protein folding offers insights into more complex questions in molecular and cellular biology as well as medicine. We also describe new approaches and tools to address complexities associated with these new areas of research. We review studies of protein stability, folding kinetics, intermediate and unfolded states, and protein self-association and aggregation.

## II. Studying protein folding

The protein folding field has witnessed significant changes and progress since the original work of Anfinsen showing that proteins can fold spontaneously [1,2]. Early *in vitro* studies showed that the folding process typically occurs on a milliseconds-to-seconds time scale, much faster than the rate estimated assuming that folding proceeds by a random search of all possible conformations. Based upon this observation, Levinthal then proposed that a random conformation search does not occur in folding and that proteins fold by specific 'folding pathways' [3]. On these pathways, the protein molecule passes through well-defined partially-structured intermediate states. Based on this view, numerous experiments and simulations were conducted to test the existence of transient folding intermediates [4,5]. It was expected that the determination of the structures and population of folding intermediates could help elucidate protein folding mechanisms. Earlier experimental studies on protein folding kinetics monitored the structural changes through relaxation of the protein's spectroscopic properties after exposing the protein to folding or unfolding conditions. The data obtained from such experiments exhibit single- or multiple-exponential time-decay: a single-exponential decay is interpreted as a signature of two-state kinetics between the native state and the denatured state, whereas models involving more than two states are required to explain multiple-exponential decay data. These experiments generally probe only the average behavior of proteins, and they are not able to provide information about the folding/unfolding process in atomic details.

The discovery of a class of simple, single-domain proteins which fold via two-state kinetics without any detectable intermediates in the early 1990s [6,7], the development of experimental techniques with improved spatial/temporal resolution[8-13], and the application of computer simulations using simplified lattice and off-lattice models[14,15] greatly enhanced our understanding of various aspects of the protein folding problem. Based on the nucleation theory [16-18], one of the early proposed mechanisms for protein folding, the nucleation-condensation model was formulated[19-21]. In this scenario, a small number of residues (folding nucleus) need to form their native contacts in order for the folding reaction to proceed fast into the native state. The cooperativity of the protein folding process is analogous to that exhibited in first-order phase transitions, which proceed via a nucleation and growth mechanism [22]. Because of these similarities, terminology used in studies of phase transitions, such as energy landscapes and nucleation, was introduced into the discussion of protein folding. The concepts of the nucleation and the free energy landscape have promoted much of the recent progress in understanding the process of protein folding. Proteins are generally thought to have evolved to exhibit globally funneled energy landscapes [23-25] which allow proteins to fold to their native states through a stochastic process in which the free energy decreases spontaneously. The unfolded state, transition state, native state and possible intermediates correspond to local minima or saddle points in the free-energy landscape.

Advances in experimental techniques such as protein engineering, nuclear magnetic resonance (NMR), mass spectrometry, hydrogen exchange, fluorescence resonance energy transfer (FRET), atomic force microscopy (AFM), have made it possible to obtain detailed information about the different conformations occurring in the folding process[26,27]. At the same time, computational methods have been developed to better interpret experimental data by using simulations to obtain structural information about the states which are populated during the

folding process. In Table 1, we list several advances in experimental and computational methodologies used for investigating the folding of model proteins.

All-atom protein models with explicit or implicit solvents were developed to study the folding thermodynamics and the unfolding dynamics of specific proteins. Technological advances in computation allowed folding simulations of small proteins and peptides at atomic detail [28-30]. However, due to the complexity and vast dimensionality of protein conformational space, all-atom MD simulations have severe limitations on the time and length scales that can be studied. Novel simulation protocols have been proposed to improve conformational sampling efficiency, including biased sampling of the free-energy surface and non-equilibrium unfolding simulations [24]. In addition, world-wide parallel computing (e.g. Folding@Home [31]) and generalized ensemble sampling techniques that involve parallel simulations of molecular systems coupled with a Monte Carlo (MC) protocol [32,33] have been successfully applied to protein folding [25,34-36].

Multi-scale modeling approaches have also been used to combine efficient conformational sampling of coarse-grained models and accuracy of all-atom models to study protein folding pathways. In this approach, iterative simulations and inter-conversion between high and low-resolution protein models are performed. Feig *et al.* developed a multi-scale modeling tool set, MMTSB [37], which integrates a simplified protein model with the MC simulation engine, MONSSTER [38], and the all-atom MD packages AMBER [39] or CHARMM[40]. Using a combination of CHARMM and discrete molecular dynamics (DMD) [41-46], Ding *et al.* reconstructed the transition state ensemble of the src-SH3 protein domain through multi-scale simulations [47]. The protein folding studies can also be facilitated by sampling protein conformations near the native state. Several native-state sampling algorithms [48,49] have been successfully utilized to study plasticity [50], cooperative interactions [51], and allostery [52] in proteins. Considering native-state ensemble naturally takes into account protein flexibility, which is shown to be crucial in structure based drug designs.

During the last five years, several tools for performing web-based analyses of protein folding dynamics have been developed. The Fold-Rate server (http://psfs.cbrc.jp/fold-rate/) [53] predicts rates of protein folding using the amino-acid sequence. The Parasol folding server (http://parasol.tamu.edu/groups/amatogroup/foldingserver) [54] predicts protein folding pathways using “probabilistic roadmaps”-based motion planning techniques. The iFold server (http://ifold.dokhlab.org) [55] allows discrete molecular dynamics (DMD) simulations of protein dynamics using simplified two-bead per residue protein models. These tools facilitate the second phase of protein folding research, whereby targeted simulations may be performed for probing the dynamics of protein folding and unfolding under controlled conditions.

DMD approaches [43-46] with simplified structural models of proteins have been extensively used for investigating general principles of protein folding and unfolding [56-60]. Dokholyan et al.[61] have highlighted the differences between molecular dynamics and DMD approaches. As opposed to the traditional MD approach of iteratively solving Newtonian equations of motion for evolving protein folding trajectory, DMD simulations solve ballistic equations of motion with square-well approximation to inter-particle interaction potentials. DMD algorithm gains efficiency over traditional MD simulations in multiple ways. First, due to ballistic modeling of particle dynamics, a larger time step can be used in DMD simulations on average, which corresponds to the time interval between fastest ballistic interactions; secondly a faster inter-particle collision detection and velocity updating algorithm is used, since only the coordinates of colliding atoms need to be updated at each collision. Additionally, faster simulation speeds are attainable with the DMD approach through simplification of protein models. Overall, an increase in simulation speed of 5–10 orders of magnitude is attainable using DMD [62]. Jang et al. [56] used DMD and simplified protein models with Gō interactions

[63] to probe protein folding kinetics. Protein folding kinetics studies using DMD simulations are reviewed in [42]. Recently, DMD simulations we used in uncovering the structural mechanisms of protein aggregation [64-66]. Among the fundamental challenges in studying protein folding using computer simulations are the time-scales and length-scales that can be investigated. DMD simulations have been shown to be useful for investigating long-timescale folding dynamics of complex biological systems such as poly-alanine aggregation [66,67] and the nucleosome core particle [68].

In addition to the extensive *in silico* and *in vitro* studies of protein folding, significant progress has been made in understanding protein folding *in vivo*. There are two major differences between protein folding *in vivo* and *in vitro*. First, protein folding *in vivo* is usually assisted by molecular machinery, such as chaperones (in an ATP-dependent manner), and often involves small molecule cofactors. Molecular chaperones such as the heat shock protein Hsp70 and chaperonin proteins facilitate protein folding, in part, by isolating the proteins from bulk cytosol [69,70]. Hartl and Horwich pioneered the research of chaperone-mediated protein folding [71], highlighting the differences between *in vivo* and *in vitro* folding mechanisms[69]. The mechanism of chaperonin GroEL mediated folding, including *in vivo* folding intermediates, has been extensively studied by Horwich and Gierasch [72,73]. Work by Landry et al. [74] showed that chaperon binding promotes α-helix formation in partially folded polypeptide chains. Horowitz et al. have investigated the role of chaperonin Cpn60-mediated hydrophobic exposure in protein folding[75,76]. Nearly one third of all proteins in living cells are coordinated to small molecule cofactors. The pioneering work of Wittung-Stafshede and coworkers on the role of cofactors in *in vivo* protein folding[77,78] demonstrated that bound metals stabilize the native fold, suggesting cofactor binding to unfolded polypeptides dramatically accelerates folding timescales[77].

A second notable difference between *in vivo* and *in vitro* protein folding is the fact that the concentrations of macromolecular solutes in cells can reach hundreds of grams per liter in cells [79], but most *in vitro* studies are performed in buffered solution with <1% of the cellular macromolecule concentration. The crowding environment *in vivo* can have a significant impact on protein stability and the native structure by changing the energy landscape of protein folding [80,81]. Dedmon *et al.*[82] showed that FlgM, a 97-residue protein from *Salmonella typhimurium* is unstructured in dilute solution, but in E.coli cells its C-terminal half is structured. McPhie *et al.*[83] found that a molten globular state of apomyoglobin at low pH is stabilized by high concentration of the inert polymer, dextran, compared to the unfolded state. Moreover, it was found that aggregate formation from human apolipoprotein C-II is significantly accelerated by the addition of dextran[84], suggesting a direct effect of molecular crowding on protein aggregation.

Over the past three decades, novel experimental techniques and simulations have yielded many significant insights in protein folding research. Important advances have been made, especially toward the understanding of folding and unfolding mechanisms, the structure of folding transition states, folding kinetics, the nature of folding pathways, and the structure of unfolded proteins and protein folding *in vivo*. Theoretical approaches to study protein folding have largely complemented experiments by providing experimentally testable hypotheses. In recent years, the rational manipulation of folding pathways and the association between protein folding and disease have marked a more applied phase of protein folding research.

## III. Protein stability

The thermodynamic stability of a protein is measured by the free energy difference between the folded state and the unfolded state ($\Delta G = G_{unfold} - G_{fold}$). It determines the fraction of folded proteins, thereby having a profound effect on protein function. Natural proteins are only

marginally stable[85]. The energetic contributions from the favorable folding forces such as hydrophobic packing, hydrogen bonding and electrostatic interactions are nearly offset by the entropic penalization of folding. As a result, the measured the $\Delta G$ values of most proteins fall in the range of 3-15 kcal/mol [86]. Due to this subtle balance between various physical interactions, a single mutation may shift the balance and significantly affect the stability of the whole protein. The accurate estimation of protein stability changes induced by mutations, measured as $\Delta\Delta G = \Delta G^{WT} - \Delta G^{Mut}$, still remains a significant challenge for computational biologists.

Experimentally, $\Delta G$ values can be obtained from denaturing experiments [6,87-90] where the protein unfolds by increasing temperature or by adding denaturing agents such as urea and guanidinium HCl (GdHCl). Theoretically, given the interactions, free energy can be obtained from statistical mechanics using the partition function, $Z$, as $G = -RT\ lnZ$. Analytical calculations of partition functions require integration over all degrees of freedom in the protein's conformational space, which is impossible in practice, except for simple models. Advances in computational biology have made possible the direct calculation of $\Delta\Delta G$ by MD or Monte Carlo simulations[28,91-94], a comprehensive review of which can be found in Ref. [95]. Using MD simulations, a $\Delta\Delta G$ value has been calculated for T157V mutant of T4 lysozyme which is in close agreement with experimental measurements [92].

The computational cost of direct $\Delta\Delta G$ estimation is still too high, however, preventing it from being applied to a large number of mutations for protein engineering. More heuristic approaches have been adapted which try to describe the free energy using empirical or effective functions taking advantage of the vast amount of known protein structures and stability measurements. Such simplifications significantly decrease the computational overhead and allow $\Delta\Delta G$ calculations of large numbers of mutants that can be compared with experimental results. In recent years, various methods have been proposed for large-scale $\Delta\Delta G$ predictions with reasonable prediction accuracies (which are commonly assessed by the linear correlation coefficient between the predicted and measured $\Delta\Delta G$ values). A comparison between these methods is listed in Table 2.

One approach utilizes statistical potentials that are developed using information from known protein structures, where the frequency distribution of amino acids conformations (such as pairwise distances and torsion angles) is used to extract effective potentials for free energy evaluations[96]. Gilis and Rooman first applied database-derived backbone dihedral potentials to study the change of thermodynamic stability upon point mutations[97-99]. They found that torsion-angle potentials predict $\Delta\Delta G$ accurately for mutations of solvent-exposed residues and that distance-dependent statistical potentials are more accurate for predicting the $\Delta\Delta G$ of buried residues. They obtained correlation coefficients of 0.55 to 0.87 for a dataset of 238 mutations. Zhou *et al.* [100] developed a knowledge-based potential using the distance-scaled finite ideal-gas reference state (DFIRE) approach and calculated $\Delta\Delta G$ for 895 mutants, which have a correlation of 0.67 with experimental measurements. Similarly, statistical potentials utilizing side-chain rotamer libraries[101], direction- and distance-dependent distributions[102], and four-body interactions[103] were also adapted for $\Delta\Delta G$ predictions and significant agreement with experimental measurements was achieved.

Another approach for large-scale $\Delta\Delta G$ predictions uses empirical functions to describe free energy changes induced by mutations and trains the parameters to recapitulate the experimental results. Guerois *et al.* [104] developed the FOLD-X energy function to study the stabilities of 1088 mutants. They used a comprehensive set of parameters to describe the van der Waals, solvation, hydrogen-bonding, electrostatic and entropic contribution to the protein stability, and obtained a correlation of 0.64 for the blind test set after training their parameters on 339 mutants[104]. Khatun *et al.* [105] utilized contact potentials to predict $\Delta\Delta G$ of three sets of

303, 658 and 1356 mutants and their prediction correlations varied between 0.45 to 0.78. Bordner and Abagyan [106] used a combination of physical energy terms, statistical energy terms and a structural descriptor with weight factors scaled to experimental data for *ΔΔG* predictions, and found a correlation of 0.59 on 908 test mutants. Saraboji *et al.* classified the available thermal denaturing data on mutations according to substitution types, secondary structures and the area of solvent accessibilities, and used the average value from each category for the prediction and obtained a correlation of 0.64 [107].

Taking advantage of the vast amount of experimental *ΔΔG* data now available, machine-learning techniques have been introduced for *ΔΔG* estimation. Capriotti *et al.* [108,109] trained a support vector machine using temperature, pH, mutations, nearby residues and relative solvent accessible area as input vectors. The support vector machine, when applied to a test set, gives a prediction correlation of 0.71. Cheng *et al.* [110] improved the support vector machine model to directly include the sequence information and obtained higher correlation prediction accuracy.

There are two significant drawbacks with training-based studies. First, improvement of the prediction accuracy relies on the available experimental stability data for parameter trainings. It is questionable whether parameters obtained from these trainings are transferable to other studies [105] since the experimentally-available mutation data may be biased (e.g., towards substitutions of large residues for small ones). Second, some mutations introduce strains in the protein backbone. To properly estimate the *ΔΔG* values, it is necessary to estimate the structural rearrangement that a protein undergoes to release the strain. To our knowledge, protein dynamics and flexibility have not been explicitly modeled in previous methods. Ignoring protein flexibility prohibits the application of current prediction methods to a wide range of mutations [100,104].

To address both these caveats, Yin *et al.* [111] developed a novel method, Eris (http://eris.dokhlab.org), for accurate and rapid evaluation of the *ΔΔG* values using the recently-developed Medusa redesign suite[112]. Eris features an all-atom force field developed from x-ray crystal structures, a fast side-chain packing algorithm, and a backbone relaxation method. The *ΔΔG* values of 595 mutants from five proteins were calculated and compared with the experimental data from the Protherm database and other sources [104,105,113,114]. Significant correlations of ≈0.75 were found between predicted and experimental *ΔΔG* values. Eris identifies and efficiently relaxes strains in the backbone, especially when clashes and backbone strains are introduced by a small-to-large amino acid substitution. Interestingly, when high-resolution structures are not available, Eris allows refinement of the backbone structure, which yields better prediction accuracy. Compared with other *ΔΔG* prediction methods, the Eris method is a unique approach that combines physical energies with efficient atomic modeling, resulting in fast and unbiased *ΔΔG* predictions.

Despite remarkable progress in the last several decades, protein stability estimation methods are still imperfect. Several obstacles must be cleared in order to achieve a more reliable method for stability estimation. Due to the complexity of sampling multi-dimensional space, the entropic free energy of a protein is difficult to evaluate and is thereby often ignored or only roughly counted in current stability estimation methods. Furthermore, since protein stability is determined by the free energy difference between the folded and unfolded states, it is crucial to model the unfolded state and its effect on protein stability (cf. section on unfolded protein states). Most importantly, the prediction of large conformational changes upon mutations remains a major challenge in protein-stability estimation.

## IV. Protein folding kinetics

To understand protein folding, important details must be taken into consideration as the protein proceeds from unfolded to native state. How fast does a protein fold? Are there multiple pathways accessed *en route* to its folded state? What are the structural characteristics that determine the path and rate of protein folding? In addition, as we realized in recent years, there is a broad class of human diseases that arises from failure of some proteins to adopt and remain in their native states, partly due to abnormal folding kinetics [115]. For example, a major cystic fibrosis-related deletion mutation of a single amino acid in the cystic fibrosis transmembrane conductance regulator (CFTR) affects its folding kinetics, but has minimal effect on its stability and structure [116]. Also, recent evidence [117] suggests that altered dynamics of superoxide dismutase (SOD1) mutants, which are non-destabilizing or even stabilizing, possibly cause the aggregation of mutants in familial amyloid sclerosis (FALS). Elucidation of protein folding kinetics has never been more important, particularly in the context of finding molecular mechanisms of protein misfolding diseases.

Probing the structural properties of intermediate states and determining folding pathways have been major experimental challenges. Nonetheless, there have been significant advances in experimental techniques. Experimental methods with different temporal resolutions allow a more detailed dissection of the folding process (Table 3). Other methods allow for investigation of protein structure (e.g. NMR, ultraviolet/visual light CD, site-directed mutagenesis, Φ-value analysis, isotope labeling) and global properties (e.g., mass spectroscopy, quasi elastic light scattering, ultracentrifugation) [118].

As experimental data on folding rates of various proteins accumulated over the years, people sought the determinants of folding rates. Plaxco and co-workers [119] showed that there is a high correlation between the folding rate and the structural properties of proteins, as defined by contact order CO, $CO = \frac{1}{LN}\sum_{N}\Delta L_{ij}$, where $L$ is the sequence length, $N$ is the total number of inter-residue atomic contacts within a cutoff distance, $\Delta L_{ij}$ is the sequence separation of contacting residues $i$ and $j$. Interestingly, assessment of other protein folding rate determinants such as local and long-range contacts were found to perform equally well as the contact order [53]. Also, contact order is a geometric property which does not take into account the distribution and strength of the interactions on the rate [120]. Fersht and coworkers found that specific interactions in the folding nucleus are equally important determinants of folding rates. For example, mutations in the folding nucleus of CI2 did not change its contact order, but result in a three order magnitude increase in the folding rate [90,121,122]. Sequence-based prediction of folding rates has also been proposed and was found to be of comparable performance to that of contact order [123,124]. Thus, there is still a debate as to whether structure-based or sequence-based prediction is a more reliable predictor of folding rates [123,124].

Extensive studies have likewise been made on the prediction of folding rates using molecular dynamics simulations. The primary bottleneck in this approach was sampling the time scales where the folding transitions are observable. Thus, early studies pioneered by Caflisch et al. [125-127] employed continuum solvent with low viscosity to observe multiple folding transitions. However, there is a nonlinear relationship between the folding time and viscosity [128], hence, the precise effect of very low viscosity on the protein folding kinetics of various systems remains unclear. To circumvent this problem, Pande et al. used “coupled ensemble dynamics” to simulate the folding of a β-hairpin from protein G using continuum solvent model [129] and united atom force-field [130] with water-like viscosity. In this and other subsequent simulations of other β-hairpins, the calculated folding rate was in close agreement with experimental measurements. A remarkable simulation in this class is a 1 μs folding simulation on the villin headpiece by Duan and Kollman [28]. Folding rate predictions of this type have

been limited to small two-state proteins. As protein size increases, it is difficult to computationally study folding kinetics. Rate predictions have been likewise performed using molecular dynamics simulations using explicit water models such as the TIP3P [131] and SPC [132] to gain additional insight into folding kinetics. Examples of MD simulations using explicit solvent which yielded experimentally consistent rates were performed by Pande et al., who observed helix-coil transitions [133] and protein folding [34]. However, a potential drawback in the use of water models is that they are parameterized to a single temperature (~298 K), and thus may bias the dynamics in non-native temperature simulations. Overall, theoretical rate determinations and their subsequent comparisons with experiments provide a test of our understanding of protein folding kinetics.

Structural investigation of the transition state ensemble (TSE) is extremely challenging experimentally, since the TSE is an unstable state whose experimental detection is very difficult. Computationally, transition state conformations may be identified using unfolding simulations (cf. section on unfolded protein states), projection into one or two reaction coordinates, validation of putative transition states through calculation of the probability to fold ($P_{fold}$), and path sampling. The intuitive appeal of low-dimensional energy landscapes in explaining simple chemical reactions inspired people to develop a similar formulation for protein folding. However, unlike simple molecules, the high degrees of freedom of a protein make the analysis formidable. Thus, several groups proposed dimensional reduction by projecting the multi-dimensional energy landscape into few relevant coordinates. The proposed reaction coordinates could be the volume of the molecule [134], the fraction of amino acids in their native conformation [135], the number of contacts between amino acids [136,137], and the fraction of native contacts in a conformation [136,137]. Some others directly tackled the transition state by developing rigorous path sampling techniques [138]. They constructed a large ensemble of transition paths, and through statistical analysis, they determined conformations whose $P_{fold}$=0.5. Although this method is computationally expensive, it is advantageous since there is no presupposed reaction coordinate.

Characteristics of the transition state ensemble have mainly been investigated by Φ-analysis [139], which involves measuring the folding kinetics and equilibrium thermodynamics of mutants containing amino acid substitutions throughout a protein. This method provides means to identify interactions mediated by specific amino acid side chains that stabilize the folding transition state[140]. Φ-analysis has been applied to a large number of proteins (such as BPTI, myoglobin, protein A, ubiquitin, SH3 domain, and the WW domain) and, recently, even to amyloidogenesis[141,142]. However, despite the prevalence of the methodology, there is debate regarding the validity and conventional interpretation of Φ-analysis, especially when the $\Delta\Delta G$ between wild type and mutant is less than 1.7 kcal/mol [2,42,143,144]. However, Fersht and coworkers argued that reliable ϕ-values can be derived from mutations in suitable proteins with $0.6 < \Delta\Delta G < 1.7$ kcal/mol[145]. Plaxco and coworkers disproved the assumed independence of the changes in free energy of transition and folded states when calculating error estimates in Φ-values [143]. They proposed a new method of error estimation that accounts for the interdependence of changes in free energy of transition and folded states.

Using simplified protein models and rapid sampling DMD, Dokholyan et al. directly observed and characterized the transition state ensemble of Src homology 3 (SH3) (Fig. 2) [47,59,146, 147]. To probe the contribution of each amino acid residue to the transition state ensemble, they calculated the Φ-values, and found high correlation between simulation and experimental Φ-values. Moreover, they also predicted that the two most kinetically important residues in folding are L24 and G64. Both L24 and G64 are experimentally-verified to be important kinetically [148].

Experiments suggest that proteins may be kinetically trapped en route to the native state [149-151], but how do proteins avoid kinetic traps? By using a Go-model scaled to include sequence-specific interactions, Khare et al. [152] found that the residues which contribute most to Cu, Zn SOD1 stability also function as “gatekeepers” that avoid kinetic traps and protein misfolding. “Gatekeeper” residues were also identified in later computational studies of the ribosomal protein S6. Stoycheva et al. [153] and Matysiak et al.[154] show that the mutations of gatekeeper residues can alter the folding landscape of S6 and shift the balance between its folding and aggregation. All these computational studies are fully consistent with experimental observations and the existence of gatekeeper residues suggests a selective pressure on avoiding misfolding in natural proteins.

The close interplay of computational and experimental efforts has advanced our knowledge of protein folding kinetics, including predicting the protein folding rate, identifying the kinetically-important residues, and characterizing the multiple pathways. For example, recent studies have demonstrated an agreement between theoretical and experimental folding free energy landscapes [154-160]. The characterization of molecular interactions responsible for different pathways opens the possibility to manipulate folding pathways. Current strategies of manipulating the folding pathway includes addition of denaturants, point mutations, and circular permutations [161]. Specifically, Kuhlman and Baker rationally engineered multiple mutations in protein L to alter its folding pathways [162]. Lowe and Itzhaki likewise recently redesigned the folding pathways of the repeat protein myotrophin [163]. Hence, one strategy to tackle kinetics-related folding abnormalities is to rationally engineer the folding pathway after the full characterization of a protein’s folding kinetics.

## V. Protein intermediate states

Proteins sample ensembles of heterogeneous conformations in solution. This emerging view of the one-to-many correspondence between protein primary sequence and its possible three-dimensional conformations also challenges the traditional paradigm that protein function is dictated by the native state. In recent years, it has been found that even for the conventionally observed small two-state proteins (~100 amino acids or less) there exist partially unfolded intermediates on the folding pathways. These intermediates are generally undetectable in kinetic folding experiments [8,116,164-168] and, therefore, are called “hidden intermediates”. In addition, mounting evidence has indicated that the intermediate states formed during protein folding and unfolding may have significant roles in protein functions (Fig. 3). The folding and unfolding intermediates impact the physiological functions of proteins by exposing cryptic post-translational modifications or ligand binding sites. The intermediates are usually weakly-populated (thermodynamic intermediates) or short-lived (kinetic intermediates), and their characterization presents a significant challenge with current experimental methods. Recent synergies between computational and experimental studies have greatly facilitated the unprecedented structural characterization of rare intermediates and suggested a functional role for these evasive conformations.

Among the traditional experimental methods, hydrogen exchange (HX)[169] is a unique tool that allows the detection and characterization of not only kinetic, but also weakly populated thermodynamic intermediates. In a typical equilibrium HX experiment, the rate at which an individual main-chain amide hydrogen exchanges with solvent deuterons is measured by NMR or mass spectrometry [170]. The exchange rates of different amide hydrogens are sensitive to local and global structural changes of proteins, and thus contain useful structural information about different protein conformations. From the amide hydrogen exchange rates measured by HX it is possible to detect and characterize weakly-populated intermediates which are inaccessible by bulk methods. Over the years, HX methods have provided considerable insight into the coarse features of intermediate state conformations for a wide variety of proteins

[169]. However, HX is limited in its ability to describe the detailed structures of intermediates, which severely restricts its applications. In contrast, computational methods can reveal structural information at much higher resolutions that cannot be accessed by experiments. Also, unlike bulk experimental methods like HX, computational methods are able to study protein motions at the single-molecule level, which enables the detection of heterogeneous conformations in an ensemble of molecules.

In recent studies, Gsponer *et al.*[171] and Dixon *et al.*[172] have developed new computational approaches to incorporate HX protection factors from NMR experiments as constraints into MD simulations to detect and/or characterize the conformations of intermediate states. Using their new approach, Gsponer *et al.*[171] were able to define the thermodynamic folding intermediates of the bacterial immunity protein Ig7. A structural comparison between this thermodynamic intermediate and kinetic intermediate determined from other methods indicates that the kinetic and thermodynamic intermediates of Ig7 are similar. In a parallel study, Dixon *et al.*[172] computationally predicted and characterized a folding intermediate of the focal adhesion targeting (FAT) domain of focal adhesion kinase (FAK), which plays critical roles in cell proliferation and migration. The detected intermediate was hypothesized to expose a cryptic phosphorylation site in the FAT domain and regulate the localization of FAK. The existence of this intermediate and the predicted structural features were later experimentally confirmed [165]. Besides HX methods, the site-specific structural information obtained from other NMR-based techniques such as relaxation dispersion spectroscopy [173], has also been incorporated into MD simulations to study weakly-populated folding intermediates [174].

A major limitation of traditional experimental methods such as HX is that they can only measure the average behavior of an ensemble of molecules and often cannot distinguish individual folding and unfolding routes or intermediates in the ensemble. By contrast, single-molecule methods enable the observation of the folding and unfolding of individual molecules and play increasingly important roles in studying intermediates[175]. Two single-molecule techniques: fluorescence resonance energy transfer [176] and force spectroscopy including optical tweezers and AFM[177], have been utilized to probe intermediate states of proteins [178-180]. Direct comparisons of results from single-molecule stretching experiments by AFM and computational simulations shed light on the mechanical unfolding of proteins and how intermediates contribute to protein function.

Using AFM, Marszalek *et al.* [179] uncovered a force-induced unfolding intermediate of Ig domain I27 from titin, a modular protein which is responsible for muscle elasticity. This experimentally-discovered intermediate was also predicted by steered-molecular dynamics (SMD) simulations [179]. Based on the results of SMD, the authors further predicted that the hydrogen-bonding network between strand A' and G in I27 mainly determines its mechanical stability. This prediction was verified in a mutagenesis study where Li *et al.* [181] showed that the point mutations on the residues participating in this hydrogen-bonding network dramatically altered the mechanical stability of I27. In a similar study of another protein, domain 10 of type III fibronectin module, which plays a pivotal role in mechanical coupling between cell surface and extracellular matrix (ECM), Gao *et al.* [182] predicted force-induced unfolding intermediates using SMD simulations. The discovered intermediates were considered to expose binding sites which are necessary for the assembly of ECM fibronectin fibrils. One of the computationally-predicted unfolding intermediates is in excellent agreement with the one observed in a more recent AFM stretching experiment [178].

The finding that folding and unfolding intermediates can have significant contributions to protein function is challenging the conventional understanding of protein structure and function that is centered on the native state. It is expected that a close synergy between computational

and experimental approaches will continue to play essential roles in characterizing these evasive protein states.

## VI. Unfolded protein states

Unveiling the structural and dynamic properties of denatured proteins is crucial for understanding the protein folding [183,184] and misfolding [185] problems. For example, the computational determination of a protein's thermodynamic stability requires an accurate approximation of the denatured state as the reference state. NMR hydrogen exchange experiments also rely on models of protein unfolded states to tabulate the intrinsic hydrogen exchange rate [186]. Furthermore, understanding the structural properties of unfolded proteins may shed light on the early events of protein folding and protein aggregation [187].

It has long been postulated that the denatured state of proteins is composed of an ensemble of featureless random coil-like conformations. According to Flory's random coil theory[188], the size of a random coil polymer, characterized by the radius of gyration $R_g$, follows a power law dependence on the length of the polymer chain, $n$, $Rg = R_0 n^{\nu}$. Here, $R_0$ is the scaling constant, which is a function of persistence length, and ν represents the power law scaling exponent. Flory predicted the exponent to be 0.6 and later a more accurate renormalization calculation obtained $\nu$=0.588 [189]. Tanford *et al.* first confirmed this random coil scaling behavior for denatured proteins [190]. Using intrinsic viscosity measurements for 12 proteins denatured by 5-6 M GuHCl, the authors obtained a scaling exponent ν=0.67±0.09. Wilkins *et al.* [191] showed that the hydrodynamic radii of sets of 8 highly denatured, disulfide-free proteins follow a power law scaling with $\nu$=0.58±0.11. Recently, Kohn *et al.* [192] reassessed the scaling behavior of denatured proteins using small angle x-ray scattering for 17 proteins of lengths varying from 8 to 549 residues. They found the scaling exponent to be $\nu$=0.598±0.029. All these experimental results confirm the random coil scaling of denatured proteins. In a random coil model of the denatured state, proteins are believed to lack persistent structure both locally and globally. The distribution of end-to-end distances or radii of gyration can be fit by a Gaussian distribution. A recent computational study [193] confirmed this behavior by generating an ensemble of protein conformations whereby only steric interactions between amino acids were considered for four different proteins. A scaling exponent of 0.58±0.02 was obtained.

The random coil scaling behavior [188,194] was originally derived for homopolymers. However, proteins are heteropolymers for which specific interactions between amino acids play an important role and determine a unique native structure. Hence, while the scaling of the sizes of denatured proteins follows the random coil scaling as shown in experiments, it does not necessarily exclude the possibility that the denatured proteins can have residual native-like structures. Under denaturing conditions, the protein can still exhibit native conformational bias and retain a certain amount of residual native structures. Mounting experimental evidence [168,195-198] supports residual native-like structural elements in the denatured state for a variety of proteins. Using residual dipolar coupling from NMR measurements, Shortle and Ackerman [195] showed that native-like topology persists under strong denaturing conditions as high as 8 M Urea for a truncated staphylococcal nuclease. By applying quasi-elastic neutron-scattering on α-lactalbumin, Bu *et al.* [196] demonstrated residual helical structure and tertiary-like interactions even in the absence of disulfide bonds and under highly denaturing conditions. Similarly, by using triple-resonance NMR, native-like topology has also been observed in protein L [197].

Several theoretical and computational studies [199-202] have addressed the role of specific interactions in conformational biasing toward the native state in the denatured states. Using a simple force field with only steric and hydrogen bond interactions, Pappu *et al.* [200]

demonstrated that denatured protein states have a strong preference for the native structure. It was suggested [199-202] that the conformational bias of native structures in the denatured state is a possible explanation of the Levinthal's paradox [3].

To reconcile the seemingly controversial properties of denatured proteins — the random coil scaling of their sizes and the presence of residual native structures — several computational works have recently been reported [203-207]. Fitzkee and Rose [206] reproduced the random coil scaling exponent using a denatured protein model with fixed secondary structure elements for a set of proteins. Tran et al. [204,207] constructed the protein denatured state ensemble at atomic resolution in the excluded volume limit. Jha et al. [205] built the denatured state using a statistical coil library. Both the Pappu and Sosnick groups found that the putative denatured state ensemble features transient local structures such as turns, strand, and helices. In the mean time, the dimension of the denatured state follows the experimentally-observed random-coil scaling exponent. Ding et al. [203] developed a computational method to model denatured proteins using a structure-based potential [63]. This interaction model is commonly used in thermodynamic and kinetics studies of protein folding [58,208-210] to model amino acid interactions. This study [203] suggested that denatured proteins follow the random coil scaling sizes *and* retain residual secondary structures akin to those observed in native protein states. Hence, these computational works provide a conceptual reconciliation between two seemingly mutually-exclusive views of protein unfolded states.

What is the physical origin of the random coil scaling of protein size along with the seemingly contradictory persistence of local structures in denatured proteins? Previous computational studies [203-205,207] suggest that the residual structures in the denatured state are limited to short-range elements, which only extend approximately seven to ten residues. The correlated fluctuation of residual structures diminishes quickly along the sequence and long-range contact formation is purely governed by the random diffusion of peptide chains. Hence, a coarse-graining process, which groups locally-interacting amino acids along the polypeptide chain into renormalized structural units (Fig. 4) reduces a denatured protein to a renormalized homopolymer. The renormalization process will result in an effective homopolymer which forms contacts due to chain diffusion. Protein sizes follow the renormalized power law scaling as proposed by Flory [188]: $Rg=R_0(N/L)^{\nu}=(R_0L^{-\nu})N^{\nu}$. Thus, we expect that the scaling exponent of denatured proteins ν is the same as for homopolymers, whose structural units are locally interacting amino acids.

The observation of residual native secondary structures in thermally-denatured protein states is consistent with a "guided-folding" scenario [211], where the rate-limiting process is the packing of the preformed secondary structures into the correct fold. In contrast, a random coil model of the denatured state without residual native-like structures implies that a protein has to overcome an excessive entropic barrier to form both the secondary and tertiary structures upon folding. The existence of persistent native-like secondary structures in the denatured state may also be responsible for the recent success of protein structure prediction using small secondary structure segments derived from the protein data bank [212]. The existence of residual native-like structures in the denatured state also provides a novel way to manipulate a protein's stability by stabilizing or destabilizing the residual structures [213].

## VII. Protein self-association and aggregation

While all the information needed for proteins to fold is encoded in their amino acid sequence [1], there are many more elements that play a part *in vivo*. In a crowded cellular environment, surrounded by interacting proteins, nascent polypeptides face a formidable challenge in finding the correct interactions that result in a folded and functional protein. Many become "trapped" in meta-stable intermediate structures which are usually recognized by proteasomal machinery

and degraded or refolded by chaperones. Alternatively, they can associate with similar misfolded proteins to form aggregates.

Extant protein sequences are the result of a long history of evolutionary refinement establishing a set of interactions defining the native state. However, the same inherent recognition that occurs between sequences within a protein is the basis for a type of self-association termed 3-dimensional domain swapping ([214], extensively catalogued in 2002 by Liu and Eisenberg [215]). Domain swapping is an important phenomenon, taking part in both normal and disease-related processes, and is intimately tied to protein folding. Domain swapping may be viewed as a natural mechanism for dealing with instability due to evolutionary changes in the amino acid sequence [216]. For example, a mutation that rigidifies a loop connecting two parts of a protein induces strain, which can be relieved without the loss of function by "swapping" the portion of the protein on one side of the loop with the corresponding part of a similar protein. Dimerization by this mechanism has advantages including a high local concentration of enzymatic activity, since two functional proteins are joined together.

Many diseases are now associated with protein aggregation and particularly with a form of ordered aggregate called the amyloid fibrils, which, regardless of the native sequence and structure of the precursor proteins, share distinct structural characteristics. From a protein folding standpoint, the inherent properties of the polypeptide chain that allow proteins with little or no sequence or structural similarity to misfold and assemble into similar high-order structures are of vital interest. Studies of aggregate structure reveal defined characteristics such as extensive hydrogen bond networks perpendicular to the fiber axis, called a cross-β conformation[217], and an α to β transition known to occur during the oligomerization of amyloid-forming proteins with significant helical content. Evidence for domain swapping as an early step in the aggregation process has been reported for several proteins [218-220]. In several aggregation-associated diseases including familial Amyotrophic Lateral Sclerosis and the transthyretin amyloidoses, the dissociation of a protein from its multimeric native state is known to be the rate limiting step for aggregation, suggesting a method for preventing aggregation by stabilizing the native interfaces of these assemblies [221-223].

Amyloid fibrils have alternately been deemed responsible for the pathologies of their various associated diseases and, more recently, credited with delaying or counteracting the observed pathologies by acting as a sink for highly cytotoxic soluble oligomers [224]. The viewpoint that soluble oligomers act as cytotoxic species has garnered widespread attention since at least 1999 when it was noted that the abundance of soluble Aβ 1-42 oligomers is inversely correlated with neuronal degeneration in Alzheimer's disease whereas amyloid levels do not correlate [225,226]. In 2003, Glabe and co-workers discovered that soluble oligomeric species from several disease-related proteins shared a common structural epitope to which an antibody was developed [227]. Later studies showed that soluble oligomers are able to disrupt the polarity of cellular membranes [144,228,229], one possible basis for disease-associated toxicity.

Various cellular protective mechanisms have evolved to ensure the proper folding of proteins. Molecular chaperones, for example, recognize misfolded proteins and provide an environment conducive to the formation of the appropriate native contacts [230]. Vast proteosomal machinery clears proteinaceous debris from the cell using ubiquitin ligases to tag misfolded proteins for degradation and removal [231,232]. One hypothesis formulated to explain the prevalence of protein aggregation in neurodegenerative diseases is that the ubiquitin proteasome loses efficiency over time causing a buildup of protein aggregates and debris in post-mitotic cells such as neurons [233]. Also, parkin, an E3 ubiquitin ligase was found to be mutated in at least half of autosomal recessive juvenile parkinsonism patients, suggesting that a deficit in the clearance of its target protein leads to an early onset of symptoms [234].

Computational studies of protein aggregation have traditionally been inadequate due to the massive complexity of the system in both time and length scale. Several approaches have been used to overcome this complexity (Fig. 5). Traditional all-atom molecular dynamics simulations have been carried out to model the aggregation of disease-related peptides such as Aβ and polyglutamine [235-237]. Other techniques seek to identify which parts within larger proteins are responsible for their aggregation behavior, resulting in the identification of sequence stretches in various proteins that are "amyloidogenic," or "hot spots" for aggregation [238-240]. Underlying this work is the idea that evolution acts to prevent aggregation by burying aggregation-prone protein sequences or otherwise prohibiting their apposition in protein structures and during folding[1]. To study the nature of subunit assembly and extension during aggregation, which are out of reach for all-atom molecular dynamics in both the size scale and time scale, simplified protein models are being utilized [65]. By accessing aggregation events that are out of the reach of experimentalists, computational studies of aggregation are an essential compliment to the experimental findings regarding aggregate structure and formation mechanisms.

Protein aggregation is now widely viewed as a fundamental property of the polypeptide chain, meaning that all of the considerations discussed in the earlier sections of this review must, to some degree, apply to the study of aggregating proteins. Therefore, the study of protein self-association and aggregation really is the study of protein folding in the context of external influences like protein concentration, localization, or evolution. As this field develops and knowledge of protein aggregation as a general phenomenon accumulates, we stand to gain not only vital tools for treating specific diseases, but also insight into the behavior of all proteins with respect to their environment.

## VIII. Conclusions

Perhaps the most significant four measures of success in the natural sciences are our abilities (i) to observe natural phenomena, (ii) to explain natural phenomena, (iii) to predict the effect of relevant variables on a specific phenomenon, and (iv) to rationally manipulate these phenomena. The protein folding field has seen significant breakthroughs according to all of these measures. Research in the protein folding field uncovered folding pathways and states that accompany the transition from unfolded to folded proteins, revealed the origin of the cooperative folding transition, allowed prediction of folding rates and changes in thermodynamic stability upon mutation, and permitted rational alteration of protein structure and folding pathways.

Given the recognition of many human maladies as "diseases of protein folding" over the past two decades, the wealth of new knowledge about the folding process is driving the study of protein folding back into its native environment, i.e. inside living organisms. The effect of the cellular environment on protein folding, e.g., how proteins fold *in vivo* and especially the behavior of "intrinsically-disordered" proteins is a highly active area of inquiry [82,242, 243],. More applied research is ongoing in the design of animal models of diseases associated with protein folding such as cystic fibrosis, ALS, Alzheimer's disease, and the prion diseases [244-246], and the design of small molecules for use in clinical trials for treating diseases like the transthyretin amyloidoses [247]. The future of the protein folding field lies in its direct application to such medical problems, and for a growing number of protein systems, the future is now.

[1]It is interesting to note here that in the case of Pmel17, which aggregates to form a "functional amyloid" involved in melanin biosynthesis, a protein seems to have evolved to aggregate at an incredible rate, perhaps to minimize the population time in a soluble oligomer form [241].

**Acknowledgements**

We thank Oana Lungu for reading the manuscript. This work was supported by the American Heart Association grant No. 0665361U, the North Carolina Biotechnology Center Grant No. 2006-MRG-1107, Cystic Fibrosis Foundation grant DOKHOL07I0, and the National Institutes of Health grants RO1GM080742-01 and R01CA084480-07.

## References

1. Anfinsen CB, Harber E, Sela M, White FH Jr. Proc Natl Acad Sci U S A 1961;47:1309–1314. [PubMed: 13683522]
2. Anfinsen CB. Science 1973;181:223–230. [PubMed: 4124164]
3. Levinthal C. J Chim Phys 1968;65:44–45.
4. Ikai A, Tanford C. Nature 1971;230:100–102. [PubMed: 4927005]
5. Tsong TY, Baldwin RL, Elson EL. Proc Natl Acad Sci U S A 1971;68:2712–2715. [PubMed: 5288248]
6. Jackson SE, Fersht AR. Biochemistry 1991;30:10428–10435. [PubMed: 1931967]
7. Jackson SE, Fersht AR. Biochemistry 1991;30:10436–10443. [PubMed: 1931968]
8. Bai Y, Sosnick TR, Mayne L, Englander SW. Science 1995;269:192–197. [PubMed: 7618079]
9. Bai Y, Englander SW. Proteins 1996;24:145–151. [PubMed: 8820481]
10. Roder H, Elove GA, Englander SW. Nature %20 1988;335:700–704.
11. Miranker A, Robinson CV, Radford SE, Aplin RT, Dobson CM. Science 1993;262:896–900. [PubMed: 8235611]
12. Jones CM, Henry ER, Hu Y, Chan CK, Luck SD, Bhuyan A, Roder H, Hofrichter J, Eaton WA. Proc Natl Acad Sci U S A 1993;90:11860–11864. [PubMed: 8265638]
13. Matouschek A, Kellis JT, Serrano L, Bycroft M, Fersht AR. Nature 1990;346:440–445. [PubMed: 2377205]
14. Mirny L, Shakhnovich E. Annu Rev Biophys Biomol Struct 2001;30:361–396. [PubMed: 11340064]
15. Shakhnovich E. Chemical Reviews 2006;106:1559–1588. [PubMed: 16683745]
16. Wetlaufer DB. Proc Natl Acad Sci U S A 1973;70:697–701. [PubMed: 4351801]
17. Wetlaufer DB. Trends Biochem Sci 1990;15:414–415. [PubMed: 2278098]
18. Go N. Annu Rev Biophys Bioeng 1983;12:183–210. [PubMed: 6347038]
19. Abkevich VI, Gutin AM, Shakhnovich EI. Biochemistry 1994;33:10026–10036. [PubMed: 8060971]
20. Fersht AR. Curr Opin Struct Biol 1997;7:3–9. [PubMed: 9032066]
21. Shakhnovich EI. Curr Opin Struct Biol 1997;7:29–40. [PubMed: 9032061]
22. Lifshits, EM.; Pitaevskii, LP. Physical Kinetics. Pergamon Press; Oxford: 1981.
23. Onuchic JN, Luthey-Schulten Z, Wolynes PG. Annu Rev Phys Chem 1997;48:545–600. [PubMed: 9348663]
24. Shea JE, Brooks CL III. Annu Rev Phys Chem 2001;52:499–535. [PubMed: 11326073]
25. Zhou RH, Berne BJ, Germain R. Proceedings of the National Academy of Sciences of the United States of America 2001;98:14931–14936. [PubMed: 11752441]
26. Fersht AR, Daggett V. Cell 2002;108:573–582. [PubMed: 11909527]
27. Dinner AR, Sali A, Smith LJ, Dobson CM, Karplus M. Trends Biochem Sci 2000;25:331–339. [PubMed: 10871884]
28. Duan Y, Kollman PA. Science 1998;282:740–744. [PubMed: 9784131]
29. Ghosh A, Elber R, Scheraga HA. Proc Natl Acad Sci U S A 2002;99:10394–10398. [PubMed: 12140363]
30. Cavalli A, Ferrara P, Caflisch A. Proteins 2002;47:305–314. [PubMed: 11948784]
31. Pande VS, Baker I, Chapman J, Elmer SP, Khaliq S, Larson SM, Rhee YM, Shirts MR, Snow CD, Sorin EJ, Zagrovic B. Biopolymers 2003;68:91–109. [PubMed: 12579582]
32. Hansmann UHE. International Journal of Quantum Chemistry 2002;90:1515–1521.
33. Hansmann UHE. Computer Physics Communications 2002;147:604–607.
34. Snow CD, Nguyen N, Pande VS, Gruebele M. Nature 2002;420:102–106. [PubMed: 12422224]
35. Gront D, Kolinski A, Skolnick J. Journal of Chemical Physics 2001;115:1569–1574.

36. Rhee YM, Pande VS. Biophysical Journal 2003;84:775–786. [PubMed: 12547762]
37. Feig M, Karanicolas J, Brooks CL. Journal of Molecular Graphics & Modelling 2004;22:377–395. [PubMed: 15099834]
38. Skolnick J, Kolinski A, Ortiz AR. Journal of Molecular Biology 1997;265:217–241. [PubMed: 9020984]
39. Pearlman DA, Case DA, Caldwell JW, Ross WS, Cheatham TE, Debolt S, Ferguson D, Seibel G, Kollman P. Computer Physics Communications 1995;91:1–41.
40. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Journal of Computational Chemistry 1983;4:187–217.
41. Zhou Y, Karplus M. Proc Natl Acad Sci U S A 1997;94:14429–14432. [PubMed: 9405629]
42. Dokholyan NV, Borreguero JM, Buldyrev SV, Ding F, Stanley HE, Shakhnovich EI. Methods Enzymol 2003;374:616–638. [PubMed: 14696390]
43. Alder BJ, Wainwright TE. Journal of Chemical Physics 1959;31:459–466.
44. Rapaport, DC. The art of molecular dynamics simulations. Cambridge University Press; Cambridge: 1997.
45. Rapaport DC. Journal of Chemical Physics 1979;71:3299–3303.
46. Bellemans A, Orban J, Vanbelle D. Molecular Physics 1980;39:781–782.
47. Ding F, Guo WH, Dokholyan NV, Shakhnovich EI, Shea JE. Journal of Molecular Biology 2005;350:1035–1050. [PubMed: 15982666]
48. Hilser VJ, Garcia-Moreno EB, Oas TG, Kapp G, Whitten ST. Chem Rev 2006;106:1545–1558. [PubMed: 16683744]
49. Ma J. Structure 2005;13:373–380. [PubMed: 15766538]
50. Cui Q, Li G, Ma J, Karplus M. J Mol Biol 2004;340:345–372. [PubMed: 15201057]
51. Liu T, Whitten ST, Hilser VJ. Proc Natl Acad Sci U S A 2007;104:4347–4352. [PubMed: 17360527]
52. Freire E. National Acad Sciences. 2000
53. Gromiha MM, Thangakani AM, Selvaraj S. Nucleic Acids Res 2006;34:W70–W74. [PubMed: 16845101]
54. Thomas S, Song G, Amato NM. Phys Biol 2005;2:S148–S155. [PubMed: 16280620]
55. Sharma S, Ding F, Nie H, Watson D, Unnithan A, Lopp J, Pozefsky D, Dokholyan NV. Bioinformatics. 2006
56. Jang H, Hall CK, Zhou Y. Biophys J 2004;86:31–49. [PubMed: 14695247]
57. Smith AV, Hall CK. Proteins 2001;44:376–391. [PubMed: 11455611]
58. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Folding & Design 1998;3:577–587. [PubMed: 9889167]
59. Borreguero JM, Dokholyan NV, Buldyrev SV, Shakhnovich EI, Stanley HE. Journal of Molecular Biology 2002;318:863–876. [PubMed: 12054829]
60. Jang H, Hall CK, Zhou Y. Biophys J 2002;83:819–835. [PubMed: 12124267]
61. Dokholyan NV. Curr Opin Struct Biol. 2006
62. Ding F, Dokholyan NV. Trends Biotechnol 2005;23:450–455. [PubMed: 16038997]
63. Go N, Abe H. Biopolymers 1981;20:991–1011. [PubMed: 7225531]
64. Urbanc B, Borreguero JM, Cruz L, Stanley HE. Methods Enzymol 2006;412:314–338. [PubMed: 17046666]
65. Peng S, Ding F, Urbanc B, Buldyrev SV, Cruz L, Stanley HE, Dokholyan NV. Phys Rev E Stat Nonlin Soft Matter Phys 2004;69:041908. [PubMed: 15169044]
66. Nguyen HD, Hall CK. Proc Natl Acad Sci U S A 2004;101:16180–16185. [PubMed: 15534217]
67. Nguyen HD, Hall CK. J Biol Chem 2005;280:9074–9082. [PubMed: 15591317]
68. Sharma S, Ding F, Dokholyan NV. Biophys J 2007;92
69. Frydman J, Hartl FU. Science 1996;272:1497–1502. [PubMed: 8633246]
70. Martin J, Hartl FU. Curr Opin Struct Biol 1997;7:41–52. [PubMed: 9032064]
71. Bukau B, Horwich AL. Cell 1998;92:351–366. [PubMed: 9476895]

72. Weissman JS, Rye HS, Fenton WA, Beechem JM, Horwich AL. Cell 1996;84:481–490. [PubMed: 8608602]
73. Feltham JL, Gierasch LM. Cell 2000;100:193–196. [PubMed: 10660042]
74. Landry SJ, Gierasch LM. Biochemistry 1991;30:7359–7362. [PubMed: 1677268]
75. Horowitz PM, Hua S, Gibbons DL. J Biol Chem 1995;270:1535–1542. [PubMed: 7829481]
76. Gorovits BM, Horowitz PM. J Biol Chem 1995;270:13057–13062. [PubMed: 7768899]
77. Wittung-Stafshede P. Acc Chem Res 2002;35:201–208. [PubMed: 11955048]
78. Leckner J, Bonander N, Wittung-Stafshede P, Malmstrom BG, Karlsson BG. Biochim Biophys Acta 1997;1342:19–27. [PubMed: 9366266]
79. Luby-Phelps K. Int Rev Cytol 2000;192:189–221. [PubMed: 10553280]
80. Minton AP. J Biol Chem 2001;276:10577–10580. [PubMed: 11279227]
81. Minton AP. J Pharm Sci 2005;94:1668–1675. [PubMed: 15986476]
82. Dedmon MM, Patel CN, Young GB, Pielak GJ. Proc Natl Acad Sci U S A 2002;99:12681–12684. [PubMed: 12271132]
83. McPhie P, Ni YS, Minton AP. J Mol Biol 2006;361:7–10. [PubMed: 16824541]
84. Hatters DM, Minton AP, Howlett GJ. J Biol Chem 2002;277:7824–7830. [PubMed: 11751863]
85. Taverna DM, Goldstein RA. Proteins 2002;46:105–109. [PubMed: 11746707]
86. Pakula AA, Sauer RT. Annual Review of Genetics 1989;23:289–310.
87. Privalov PL. Adv Protein Chem 1979;33:167–241. [PubMed: 44431]
88. Becktel WJ, Schellman JA. Biopolymers 1987;26:1859–1877. [PubMed: 3689874]
89. Makhatadze GI, Privalov PL. J Mol Biol 1992;226:491–505. [PubMed: 1322462]
90. Jackson SE, Moracci M, elMasry N, Johnson CM, Fersht AR. Biochemistry 1993;32:11259–11269. [PubMed: 8218191]
91. Bash PA, Singh UC, Langridge R, Kollman PA. Science 1987;236:564–568. [PubMed: 3576184]
92. Dang LX, Merz KM, Kollman PA. Journal of the American Chemical Society 1989;111:8505–8508.
93. Kollman PA, Massova I, Reyes C, Kuhn B, Huo SH, Chong L, Lee M, Lee T, Duan Y, Wang W, Donini O, Cieplak P, Srinivasan J, Case DA, Cheatham TE. Accounts of Chemical Research 2000;33:889–897. [PubMed: 11123888]
94. Vorobjev YN, Hermans J. Biophysical Chemistry 1999;78:195–205. [PubMed: 10343388]
95. Kollman P. Chemical Reviews 1993;93:2395–2417.
96. Lazaridis T, Karplus M. Curr Opin Struct Biol 2000;10:139–145. [PubMed: 10753811]
97. Gilis D, Rooman M. Journal of Molecular Biology 1996;257:1112–1126. [PubMed: 8632471]
98. Gilis D, Rooman M. Journal of Molecular Biology 1997;272:276–290. [PubMed: 9299354]
99. Gilis D, Rooman M. Protein Engineering 2000;13:849–856. [PubMed: 11239084]
100. Zhou HY, Zhou YQ. Protein Sci 2003;12:2121.
101. Ota M, Isogai Y, Nishikawa K. Protein Engineering 2001;14:557–564. [PubMed: 11579224]
102. Hoppe C, Schomburg D. Protein Sci 2005;14:2682–2692. [PubMed: 16155198]
103. Carter CW, LeFebvre BC, Cammer SA, Tropsha A, Edgell MH. J Mol Biol 2001;311:625–638. [PubMed: 11518520]
104. Guerois R, Nielsen JE, Serrano L. Journal of Molecular Biology 2002;320:369–387. [PubMed: 12079393]
105. Khatun J, Khare SD, Dokholyan NV. J Mol Biol 2004;336:1223–1238. [PubMed: 15037081]
106. Bordner AJ, Abagyan RA. Proteins-Structure Function and Bioinformatics 2004;57:400–413.
107. Saraboji K, Gromiha MM, Ponnuswamy MN. Biopolymers 2006;82:80–92. [PubMed: 16453276]
108. Capriotti E, Fariselli P, Calabrese R, Casadio R. Bioinformatics 2005;21:54–58.
109. Kuhlman B, Baker D. Current Opinion in Structural Biology 2004;14:89–95. [PubMed: 15102454]
110. Cheng JL, Randall A, Baldi P. Proteins-Structure Function and Bioinformatics 2006;62:1125–1132.
111. Yin SY, Ding F, Dokholyan NV. Nature Methods 2007;4:2.
112. Ding F, Dokholyan NV. Public Library of Science Computational Biology 2006;2:e85. [PubMed: 16839198]

113. Edgell MH, Sims DA, Pielak GJ, Yi F. Biochemistry 2003;42:7587–7593. [PubMed: 12809515]
114. Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A. Nucleic Acids Research 2004;32:D120–D121. [PubMed: 14681373]
115. Chiti F, Dobson CM. Annual Review of Biochemistry 2006;75:333–366.
116. Thibodeau PH, Brautigam CA, Machius M, Thomas PJ. Nature Structural & Molecular Biology 2005;12:10–16.
117. Khare SD, Dokholyan NV. Proceedings of the National Academy of Sciences of the United States of America 2006;103:3147–3152. [PubMed: 16488975]
118. Nolting, B. Protein Folding Kinetics. Springer-Verlag; Berlin, Germany: 2006.
119. Plaxco KW, Simons KT, Baker D. Journal of Molecular Biology 1998;277:985–994. [PubMed: 9545386]
120. Munoz V, Eaton WA. PNAS 1999;96:11311–11316. [PubMed: 10500173]
121. Otzen DE, Itzhaki LS, elMasry NF, Jackson SE, Fersht AR. Proc Natl Acad Sci U S A 1994;91:10422–10425. [PubMed: 7937967]
122. Li A, Daggett V. Proc Natl Acad Sci U S A 1994;91:10430–10434. [PubMed: 7937969]
123. Ivankov DN, Finkelstein AV. Proceedings of the National Academy of Sciences of the United States of America 2004;101:8942–8944. [PubMed: 15184682]
124. Kuznetsov IB, Rackovsky S. Proteins-Structure Function and Genetics 2004;54:333–341.
125. Ferrara P, Apostolakis J, Caflisch A. J Phys Chem B 2000;104:5000–5010.
126. Ferrara P, Caflisch A. Proc Natl Acad Sci U S A 2000;97:10780–10785. [PubMed: 10984515]
127. Hiltpold A, Ferrara P, Gsponer J, Caflisch A. J Phys Chem B 2000;104:10080–10086.
128. Zagrovic B, Pande V. Journal of Computational Chemistry 2003;24:1432–1436. [PubMed: 12868108]
129. Qiu D, Shenkin PS, Hollinger FP, Still WC. J Phys Chem A 1997;101:3005–3014.
130. Jorgensen WL, Tiradorives J. Journal of the American Chemical Society 1988;110:1657–1666.
131. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. The Journal of Chemical Physics 1983;79:926–935.
132. Berendsen, H.; Potsma, J.; van Gunsteren, W.; Hermans, J. Intermolecular Forces. Dordrecht: Reidel; 1981. p. 331-334.
133. Sorin EJ, Pande VS. Biophysical Journal 2005;88:2472–2493. [PubMed: 15665128]
134. Shakhnovich EI, Finkelstein AV. Biopolymers 1989;28:1667–1680. [PubMed: 2597723]
135. Bryngelson JD, Wolynes PG. Proc Natl Acad Sci U S A 1987;84:7524–7528. [PubMed: 3478708]
136. Shakhnovich E, Farztdinov G, Gutin AM, Karplus M. Phys Rev Lett 1991;67:1665. [PubMed: 10044213]
137. Sali A, Shakhnovich E, Karplus M. Nature 1994;369:248–251. [PubMed: 7710478]
138. Bolhuis PG, Chandler D, Dellago C, Geissler PL. Annual Review of Physical Chemistry 2002;53:291–318.
139. Matouschek A, Kellis JT, Serrano L, Fersht AR. Nature 1989;340:122–126. [PubMed: 2739734]
140. Fersht AR, Matouschek A, Serrano L. Journal of Molecular Biology 1992;224:771–782. [PubMed: 1569556]
141. Day R, Daggett V. Protein Simulations 2003;66:373–403.
142. Scheraga HA. Febs Journal 2005;272:359.
143. Ruczinski I, Sosnick TR, Plaxco KW. Protein Sci 2006;15:2257–2264. [PubMed: 17008714]
144. Baglioni S, Casamenti F, Bucciantini M, Luheshi LM, Taddei N, Chiti F, Dobson CM, Stefani M. J Neurosci 2006;26:8160–8167. [PubMed: 16885229]
145. Fersht AR, Sato S. PNAS 2004;101:7976–7981. [PubMed: 15150406]
146. Ding F, Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Biophysical Journal 2002;83:3525–3532. [PubMed: 12496119]
147. Ding F, Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Journal of Molecular Biology 2002;324:851–857. [PubMed: 12460582]

148. Di Nardo AA, Korzhnev DM, Stogios PJ, Zarrine-Afsar A, Kay LE, Davidson AR. Proceedings of the National Academy of Sciences of the United States of America 2004;101:7954–7959. [PubMed: 15148398]
149. Chen YR, Clark AC. Protein Science 2004;13:2196–2206. [PubMed: 15273313]
150. Chen YR, Clark AC. Biochemistry 2003;42:6310–6320. [PubMed: 12755636]
151. Reader JS, Van Nuland NAJ, Thompson GS, Ferguson SJ, Dobson CM, Radford SE. Protein Sci 2001;10:1216–1224. [PubMed: 11369860]
152. Khare SD, Ding F, Dokholyan NV. J Mol Biol 2003;334:515–525. [PubMed: 14623191]
153. Stoycheva AD, Brooks CL III, Onuchic JN. J Mol Biol 2004;340:571–585. [PubMed: 15210355]
154. Matysiak S, Clementi C. J Mol Biol 2006;363:297–308. [PubMed: 16959265]
155. Shea JE, Onuchic JN, Brooks CL III. Proc Natl Acad Sci U S A 2002;99:16064–16068. [PubMed: 12446834]
156. Nguyen H, Jager M, Moretto A, Gruebele M, Kelly JW. Proceedings of the National Academy of Sciences of the United States of America 2003;100:3948–3953. [PubMed: 12651955]
157. Karanicolas J, Brooks CL. Proceedings of the National Academy of Sciences of the United States of America 2003;100:3954–3959. [PubMed: 12655041]
158. Fernandez-Escamilla AM, Cheung MS, Vega MC, Wilmanns M, Onuchic JN, Serrano L. Proceedings of the National Academy of Sciences of the United States of America 2004;101:2834–2839. [PubMed: 14978284]
159. Chavez LL, Onuchic JN, Clementi C. Journal of the American Chemical Society 2004;126:8426–8432. [PubMed: 15237999]
160. Onuchic JN, Wolynes PG. Current Opinion in Structural Biology 2004;14:70–75. [PubMed: 15102452]
161. Lindberg M, Oliveberg M. Current Opinion in Structural Biology 2007;17:21–29. [PubMed: 17251003]
162. Nauli S, Kuhlman B, Baker D. Nature Structural Biology 2001;8:602–605.
163. Lowe AR, Itzhaki LS. PNAS 2007;104:2679–2684. [PubMed: 17299057]
164. Feng HQ, Vu ND, Bai YW. Journal of Molecular Biology 2005;346:345–353. [PubMed: 15663949]
165. Zhou Z, Feng H, Bai Y. Proteins 2006;65:259–265. [PubMed: 16909417]
166. Fersht AR. Proceedings of the National Academy of Sciences of the United States of America 2000;97:1525–1529. [PubMed: 10677494]
167. Kato H, Feng HQ, Bai YW. Journal of Molecular Biology 2007;365:870–880. [PubMed: 17109883]
168. Religa TL, Markson JS, Mayor U, Freund SM, Fersht AR. Nature 2005;437:1053–1056. [PubMed: 16222301]
169. Englander SW. Annu Rev Biophys Biomol Struct 2000;29:213–238. [PubMed: 10940248]
170. Englander SW. J Am Soc Mass Spectrom 2006;17:1481–1489. [PubMed: 16876429]
171. Gsponer J, Hopearuoho H, Whittaker SB, Spence GR, Moore GR, Paci E, Radford SE, Vendruscolo M. Proc Natl Acad Sci U S A 2006;103:99–104. [PubMed: 16371468]
172. Dixon RD, Chen Y, Ding F, Khare SD, Prutzman KC, Schaller MD, Campbell SL, Dokholyan NV. Structure 2004;12:2161–2171. [PubMed: 15576030]
173. Mittermaier A, Kay LE. Science 2006;312:224–228. [PubMed: 16614210]
174. Korzhnev DM, Salvatella X, Vendruscolo M, Di Nardo AA, Davidson AR, Dobson CM, Kay LE. Nature 2004;430:586–590. [PubMed: 15282609]
175. Zhuang X, Rief M. Curr Opin Struct Biol 2003;13:88–97. [PubMed: 12581665]
176. Selvin PR. Nat Struct Biol 2000;7:730–734. [PubMed: 10966639]
177. Clausen-Schaumann H, Seitz M, Krautbauer R, Gaub HE. Curr Opin Chem Biol 2000;4:524–530. [PubMed: 11006539]
178. Li L, Huang HH, Badilla CL, Fernandez JM. J Mol Biol 2005;345:817–826. [PubMed: 15588828]
179. Marszalek PE, Lu H, Li H, Carrion-Vazquez M, Oberhauser AF, Schulten K, Fernandez JM. Nature 1999;402:100–103. [PubMed: 10573426]
180. Cecconi C, Shank EA, Bustamante C, Marqusee S. Science 2005;309:2057–2060. [PubMed: 16179479]

181. Li H, Carrion-Vazquez M, Oberhauser AF, Marszalek PE, Fernandez JM. Nat Struct Biol 2000;7:1117–1120. [PubMed: 11101892]
182. Gao M, Craig D, Lequin O, Campbell ID, Vogel V, Schulten K. Proc Natl Acad Sci U S A 2003;100:14784–14789. [PubMed: 14657397]
183. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Proteins 1995;21:167–195. [PubMed: 7784423]
184. Baldwin RL. Adv Protein Chem 2002;62:361–367. [PubMed: 12418110]
185. Dobson CM. Semin Cell Dev Biol 2004;15:3–16. [PubMed: 15036202]
186. Bai Y, Milne JS, Mayne L, Englander SW. Proteins 1993;17:75–86. [PubMed: 8234246]
187. Uversky VN, Fink AL. Biochim Biophys Acta 2004;1698:131–153. [PubMed: 15134647]
188. Flory, PJ. Principles of Polymer Chemistry. Cornell University Press; Ithaca, NY: 1965.
189. Le Guillou JC, Zinn-Justin J. Phys Rev Lett 1977;39:95–98.
190. Tanford C, Kawahara K, Lapanje S, Hooker TM Jr, Zarlengo MH, Salahuddin A, Aune KC, Takagi T. J Am Chem Soc 1967;89:5023–5029. [PubMed: 6074805]
191. Wilkins DK, Grimshaw SB, Receveur V, Dobson CM, Jones JA, Smith LJ. Biochemistry 1999;38:16424–16431. [PubMed: 10600103]
192. Kohn JE, Millett IS, Jacob J, Zagrovic B, Dillon TM, Cingel N, Dothager RS, Seifert S, Thiyagarajan P, Sosnick TR, Hasan MZ, Pande VS, Ruczinski I, Doniach S, Plaxco KW. Proc Natl Acad Sci U S A 2004;101:12491–12496. [PubMed: 15314214]
193. Goldenberg DP. J Mol Biol 2003;326:1615–1633. [PubMed: 12595269]
194. de Gennes, PG. Scaling Concepts in Polymer Physics. Cornell University Press; Ithaca, NY: 1979.
195. Shortle D, Ackerman MS. Science 2001;293:487–489. [PubMed: 11463915]
196. Bu Z, Cook J, Callaway DJ. J Mol Biol 2001;312:865–873. [PubMed: 11575938]
197. Yi Q, Scalley-Kim ML, Alm EJ, Baker D. Journal of Molecular Biology 2000;299:1341–1351. [PubMed: 10873457]
198. Platt GW, McParland VJ, Kalverda AP, Homans SW, Radford SE. J Mol Biol 2005;346:279–294. [PubMed: 15663944]
199. Zwanzig R, Szabo A, Bagchi B. Proc Natl Acad Sci U S A 1992;89:20–22. [PubMed: 1729690]
200. Pappu RV, Srinivasan R, Rose GD. Proc Natl Acad Sci U S A 2000;97:12565–12570. [PubMed: 11070081]
201. Srinivasan R, Rose GD. Biophys Chem 2002;101-102:167–171. [PubMed: 12487998]
202. Fitzkee NC, Rose GD. Protein Sci 2004;13:633–639. [PubMed: 14767081]
203. Ding F, Jha RK, Dokholyan NV. Structure 2005;13:1047–1054. [PubMed: 16004876]
204. Tran HT, Wang X, Pappu RV. Biochemistry 2005;44:11369–11380. [PubMed: 16114874]
205. Jha AK, Colubri A, Freed KF, Sosnick TR. Proc Natl Acad Sci U S A 2005;102:13099–13104. [PubMed: 16131545]
206. Fitzkee NC, Rose GD. Proc Natl Acad Sci U S A 2004;101:12497–12502. [PubMed: 15314216]
207. Tran HT, Pappu RV. Biophys J 2006;91:1868–1886. [PubMed: 16766618]
208. Clementi C, Nymeyer H, Onuchic JN. Journal of Molecular Biology 2000;298:937–953. [PubMed: 10801360]
209. Yang S, Cho SS, Levy Y, Cheung MS, Levine H, Wolynes PG, Onuchic JN. Proc Natl Acad Sci U S A 2004;101:13786–13791. [PubMed: 15361578]
210. Dokholyan NV, Buldyrev SV, Stanley HE, Shakhnovich EI. Journal of Molecular Biology 2000;296:1183–1188. [PubMed: 10698625]
211. Baldwin RL, Rose GD. Trends in Biochemical Sciences 1999;24:77–83. [PubMed: 10098403]
212. Bradley P, Chivian D, Meiler J, Misura KM, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss CE, Baker D. Proteins 2003;53(Suppl 6):457–468. [PubMed: 14579334]
213. Bowler BE. Mol Biosyst 2007;3:88–99. [PubMed: 17245488]
214. Bennett MJ, Choe S, Eisenberg D. Proc Natl Acad Sci U S A 1994;91:3127–3131. [PubMed: 8159715]

215. Liu Y, Eisenberg D. Protein Sci 2002;11:1285–1299. [PubMed: 12021428]
216. Schymkowitz JW, Rousseau F, Wilkinson HR, Friedler A, Itzhaki LS. Nat Struct Biol 2001;8:888–892. [PubMed: 11573096]
217. Blake C, Serpell L. Structure 1996;4:989–998. [PubMed: 8805583]
218. Eakin CM, Attenello FJ, Morgan CJ, Miranker AD. Biochemistry 2004;43:7808–7815. [PubMed: 15196023]
219. Knaus KJ, Morillas M, Swietnicki W, Malone M, Surewicz WK, Yee VC. Nat Struct Biol 2001;8:770–774. [PubMed: 11524679]
220. Sanders A, Jeremy CC, Higgins LD, Giannini S, Conroy MJ, Hounslow AM, Waltho JP, Staniforth RA. J Mol Biol 2004;336:165–178. [PubMed: 14741212]
221. Johnson SM, Wiseman RL, Sekijima Y, Green NS, mski-Werner SL, Kelly JW. Acc Chem Res 2005;38:911–921. [PubMed: 16359163]
222. Khare SD, Caplow M, Dokholyan NV. Proc Natl Acad Sci U S A 2004;101:15094–15099. [PubMed: 15475574]
223. Ray SS, Nowak RJ, Strokovich K, Brown RH, Walz T, Lansbury PT. Biochemistry 2004;43:4899–4905. [PubMed: 15109247]
224. Rao KS, Hegde ML, Anitha S, Musicco M, Zucca FA, Turro NJ, Zecca L. Prog Neurobiol 2006;78:364–373. [PubMed: 16682109]
225. Lue LF, Kuo YM, Roher AE, Brachova L, Shen Y, Sue L, Beach T, Kurth JH, Rydel RE, Rogers J. Am J Pathol 1999;155:853–862. [PubMed: 10487842]
226. McLean CA, Cherny RA, Fraser FW, Fuller SJ, Smith MJ, Beyreuther K, Bush AI, Masters CL. Ann Neurol 1999;46:860–866. [PubMed: 10589538]
227. Kayed R, Head E, Thompson JL, McIntire TM, Milton SC, Cotman CW, Glabe CG. Science 2003;300:486–489. [PubMed: 12702875]
228. Demuro A, Mina E, Kayed R, Milton SC, Parker I, Glabe CG. J Biol Chem 2005;280:17294–17300. [PubMed: 15722360]
229. Kayed R, Sokolov Y, Edmonds B, McIntire TM, Milton SC, Hall JE, Glabe CG. J Biol Chem 2004;279:46363–46366. [PubMed: 15385542]
230. Bukau B, Weissman J, Horwich A. Cell 2006;125:443–451. [PubMed: 16678092]
231. Berke SJ, Paulson HL. Curr Opin Genet Dev 2003;13:253–261. [PubMed: 12787787]
232. Rubinsztein DC. Nature 2006;443:780–786. [PubMed: 17051204]
233. Keller JN, Gee J, Ding Q. Ageing Res Rev 2002;1:279–293. [PubMed: 12039443]
234. Lucking CB, Durr A, Bonifati V, Vaughan J, De MG, Gasser T, Harhangi BS, Meco G, Denefle P, Wood NW, Agid Y, Brice A. N Engl J Med 2000;342:1560–1567. [PubMed: 10824074]
235. Han W, Wu YD. J Am Chem Soc 2005;127:15408–15416. [PubMed: 16262404]
236. Khare SD, Ding F, Gwanmesia KN, Dokholyan NV. PLoS Comput Biol 2005;1:230–235. [PubMed: 16158094]
237. Marchut AJ, Hall CK. Comput Biol Chem 2006;30:215–218. [PubMed: 16678490]
238. Fernandez-Escamilla AM, Rousseau F, Schymkowitz J, Serrano L. Nat Biotechnol 2004;22:1302–1306. [PubMed: 15361882]
239. Khare SD, Wilcox KC, Gong P, Dokholyan NV. Proteins 2005;61:617–632. [PubMed: 16152647]
240. Thompson MJ, Sievers SA, Karanicolas J, Ivanova MI, Baker D, Eisenberg D. Proc Natl Acad Sci U S A 2006;103:4074–4078. [PubMed: 16537487]
241. Fowler DM, Koulov AV, ory-Jost C, Marks MS, Balch WE, Kelly JW. PLoS Biol 2006;4:e6. [PubMed: 16300414]
242. Bryant JE, Lecomte JT, Lee AL, Young GB, Pielak GJ. Biochemistry 2005;44:9275–9279. [PubMed: 15981993]
243. Gething MJ, Sambrook J. Nature 1992;355:33–45. [PubMed: 1731198]
244. Hensley K, Mhatre M, Mou S, Pye QN, Stewart C, West M, Williamson KS. Antioxid Redox Signal 2006;8:2075–2087. [PubMed: 17034351]
245. Games D, Buttini M, Kobayashi D, Schenk D, Seubert P. J Alzheimers Dis 2006;9:133–149. [PubMed: 16914852]

246. Harris DA, Chiesa R, Drisaldi B, Quaglio E, Migheli A, Piccardo P, Ghetti B. Clin Lab Med 2003;23:175–186. [PubMed: 12733431]

247. Sekijima Y, Dendle MA, Kelly JW. Amyloid 2006;13:236–249. [PubMed: 17107884]

248. Wildegger G, Kiefhaber T. J Mol Biol 1997;270:294–304. [PubMed: 9236130]

249. Itzhaki LS, Otzen DE, Fersht AR. J Mol Biol 1995;254:260–288. [PubMed: 7490748]

250. Riddle DS, Grantcharova VP, Santiago JV, Alm E, Ruczinski I, Baker D. Nat Struct Biol 1999;6:1016–1024. [PubMed: 10542092]

251. Grantcharova VP, Riddle DS, Santiago JV, Baker D. Nat Struct Biol 1998;5:714–720. [PubMed: 9699636]

252. Daggett V, Li A, Itzhaki LS, Otzen DE, Fersht AR. J Mol Biol 1996;257:430–440. [PubMed: 8609634]

253. Du R, Pande VS, Grosberg AY, Tanaka T, Shakhnovich ES. Journal of Chemical Physics 1998;108:334–350.

254. Grantcharova VP, Baker D. Biochemistry 1997;36:15685–15692. [PubMed: 9398297]

255. Dokholyan NV, Li L, Ding F, Shakhnovich EI. Proceedings of the National Academy of Sciences of the United States of America 2002;99:8637–8641. [PubMed: 12084924]

256. Northey JGB, Di Nardo AA, Davidson AR. Nature Structural Biology 2002;9:126–130.

257. Lazaridis T, Karplus M. Science 1997;278:1928–1931. [PubMed: 9395391]

258. Borreguero JM, Ding F, Buldyrev SV, Stanley HE, Dokholyan NV. Biophysical Journal 2004;87:521–533. [PubMed: 15240485]

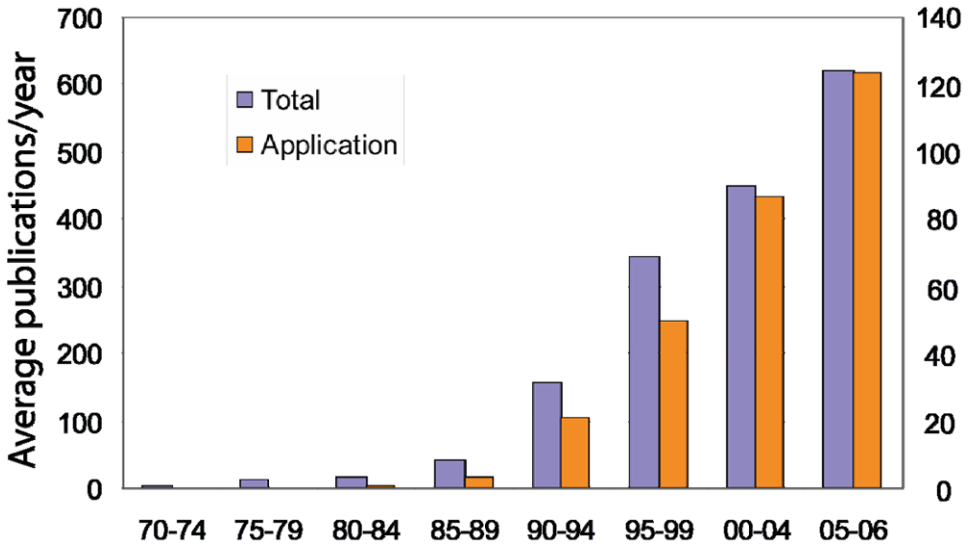


**Figure 1. Growth of the Protein Folding Field**

The average number of publications per year in protein folding field (left y axis) and the average number of publications per year that are dedicated to application (right y axis) were plotted every five years between 1970 and 2004, and 2005-2006. The first dataset was generated by searching articles in PubMed that contain the keyword ‘protein folding’ or ‘protein unfolding’ in either title or abstract. The second dataset was extracted from the previous dataset by searching with the following additional keywords: ‘engineering’, ‘design’, ‘misfolding’, ‘aggregation’, ’amyloid’ and ‘amyloid disease’.

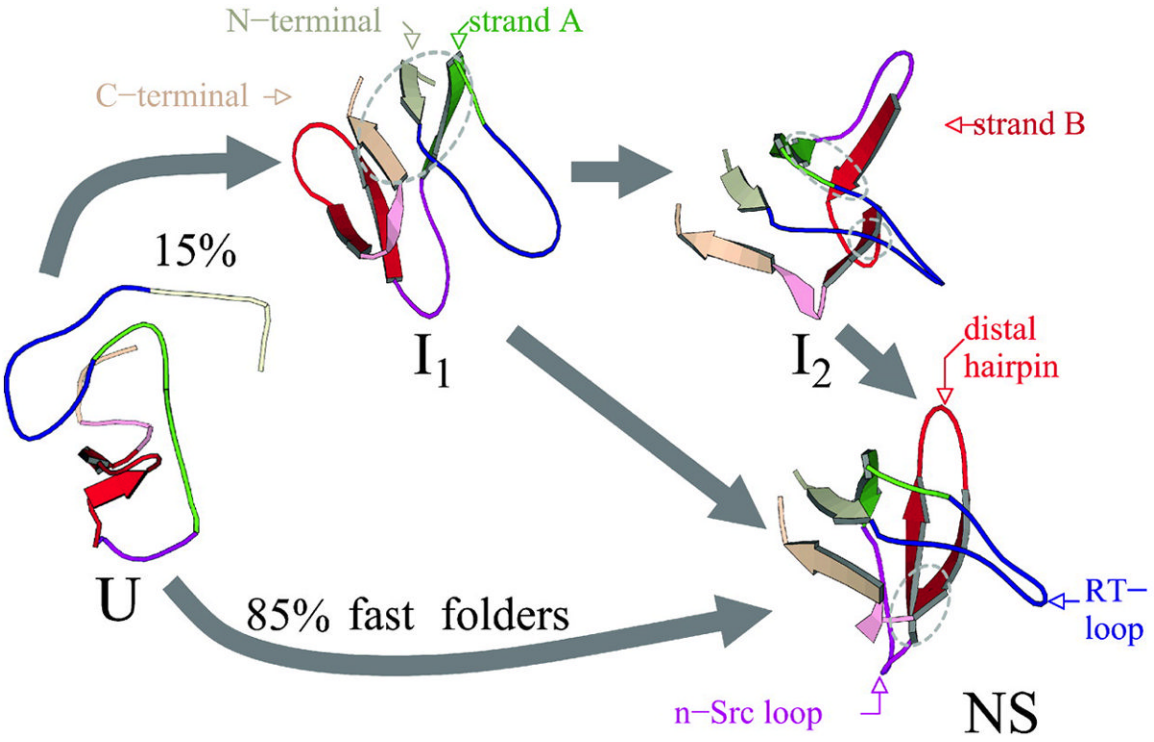


**Figure 2. Fast and slow folding pathways**

Molecular dynamics simulations of the c-Crk SH3 folding show multiple folding pathways via only one or two intermediates. [Reprinted with permission from ref. [258]. Copyright Biophysical Society.]

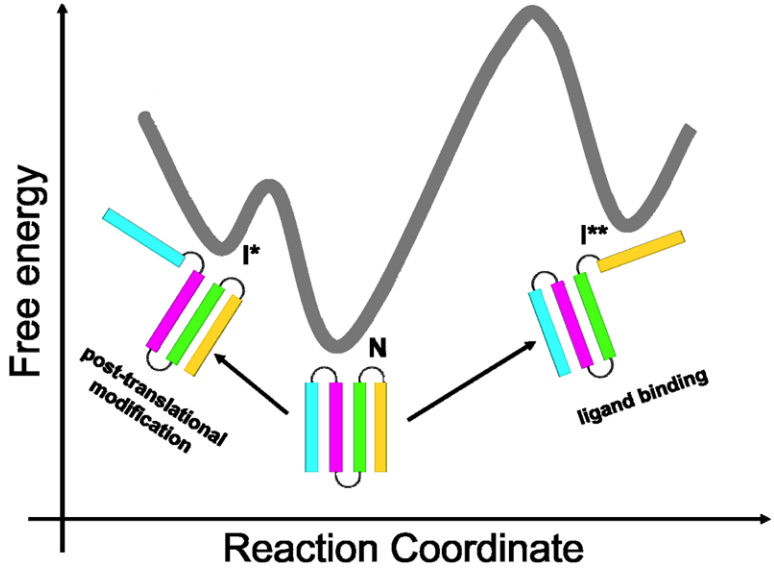


**Figure 3. Protein intermediate states**

Intermediate states are meta-stable in protein free energy landscapes (N: the native state; I*, I**: intermediates). They may play a significant role in protein function, including exposing cryptic posttranslational modifications or ligand binding sites.

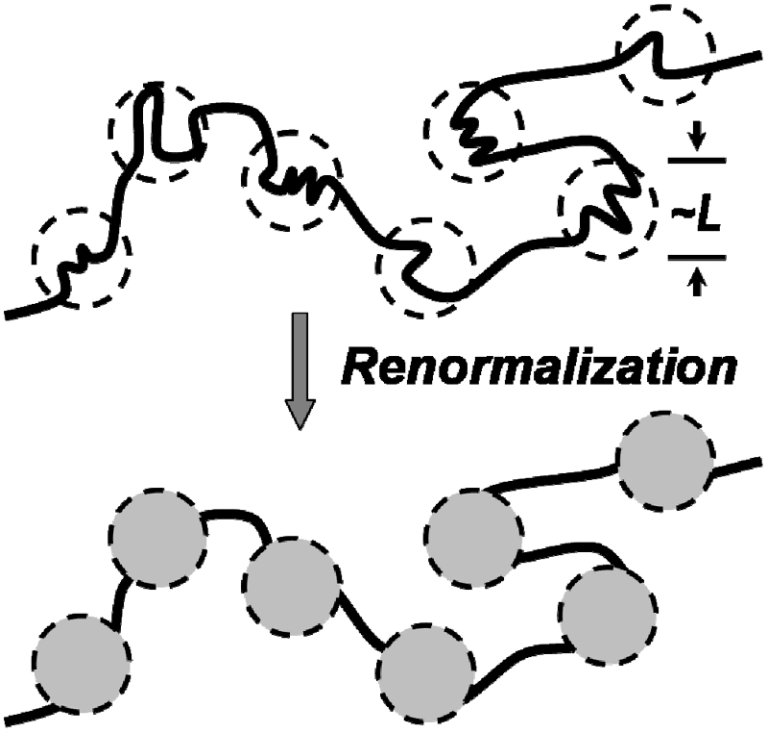


**Figure 4. Unfolded protein states**

The unfolded state of a protein features residual local structures, which span a short segment of approximately 10 residues (L~10). This structural correlation quickly decays as the segment length increases. A renormalization process, which groups the local amino acid residues into a coarse-grained bead, reduces the unfolded protein into an effective non-interacting polymer.

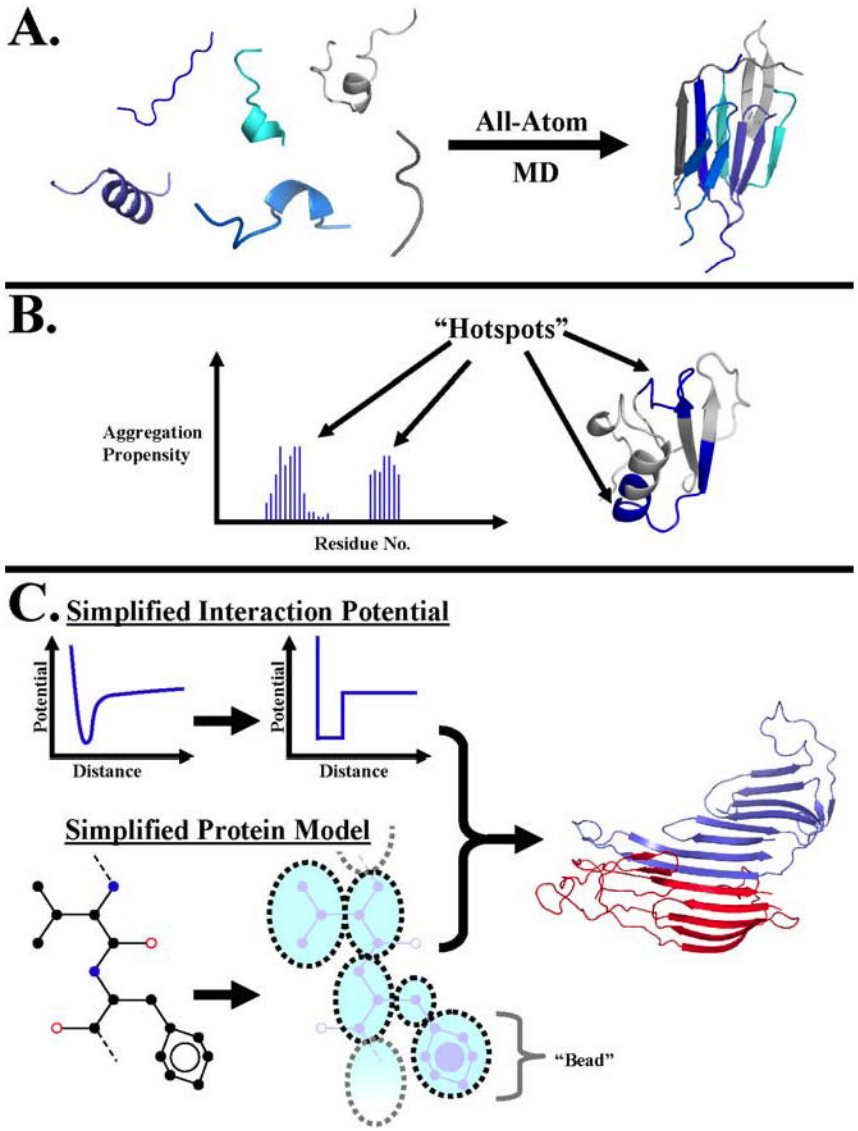


**Figure 5. Computational studies of protein self-association and aggregation**

Different computational techniques have been utilized to study various aspects of protein self-association and aggregation. (A) All-atom molecular dynamics is used to model the aggregation of short peptides. (B) Simulations of peptides from within larger proteins are used to suggest aggregation "hotspots." (C) By combining simplified interaction models and protein models, aggregating systems that are inaccessible by traditional molecular dynamics due to size and time limitations can be studied. The curves show the approximation of a continuous interaction potential by a square well as used in DMD [58]. The dotted circles represent "beads" in the model which take the place of several atoms in the original protein. The structure to the right shows the self-association of two identical proteins forming an extended β-sheet.

## Table 1

**Protein folding in select model systems**

A sampling of experimental and theoretical approaches for probing protein folding for three model protein systems: Chymotrypsin Inhibitor 2 (CI2), Lysozyme, and src/Fyn SH3 Domain.

| Theme | Relevant experiments and theoretical models | Contribution to Protein Folding Research |
|---|---|---|
| Mechanism of folding/ unfolding | Equillibrium denaturation by guanidinium chloride[6] | CI2 unfolding and refolding follows two-state transition. Lysozyme folding intermediates obstruct formation of transition state, but does not change the folding rate [248]. |
| | Protein engineering, Φ-value analysis [249]. | CI2 folding supports nucleation-condensation model [19,20]. CI2 transition state has secondary, tertiary structure elements [121]. Native topology and hydrogen bonds mediates SH3 folding [250,251]. |
| | Quantitative Φ–value analysis using MD simulations [122]. | Postulated structure of CI2 unfolding transition state. |
| | Free-energy landscape: protein folding funnel [23]. | Statistical description of protein folding process. Role of water in facilitating protein folding [155]. |
| Transition State Structure | MD simulations of CI2 transition state [252]. | Folding of CI2 is cooperative. |
| | Multiscalar modeling and DMD simulations. | Identification of src SH3 residues critical to folding nucleus [47]. |
| | Monte Carlo simulations on lattice models [253] | $p_{fold}$ as a reaction coordinate for protein folding. |
| Folding Kinetics | Equillibrium and stopped-flow fluorescence [254]. | Src SH3 unfolding is cooperative; its denatured state may be compact under native conditions [254]. |
| | Relaxation dispersion NMR of Fyn SH3 [174]. | Identified and characterized low-population folding intermediates |
| | Graph representation of CI2 and src-SH3 conformation [255]. | Protein network contact topology determines proteins' ability to fold. |
| | Φ-value analysis of SH3 [256]. | Hydrophobic core composition is another determinant of protein folding rate |
| | Multiple MD simulations of CI2 unfolding [257]. | Preferred pathway for protein folding on a funnel-like average energy surface. |
| Unfolded proteins structure | Unfolded proteins' NMR [195]. Size of measurements of various unfolded proteins [192]. | Denatured proteins have a strong local conformational bias towards native state. On the other hand, the scaling of protein sizes in the unfolded state suggests a random-coil like conformations. |
| | Computational models of denatured proteins [192,203]. | Unfolded states features local native-like structures (short-rage correlations), but the correlations decays quickly. Protein behaves as a "renormalized" random coil after grouping local structures together. |

Table 2

**Comparison of different ΔΔG calculation approaches**

The approaches are evaluated base on the calculation speed, dependence on parameter training using existing stability data, side-chain and backbone flexibility modeling capability, and transferability to study other protein properties.

| Methods | Speed | Parameter Training | Side-chain Flexibility | Backbone Flexibility | Transferability |
|---|---|---|---|---|---|
| Ab initio Simulation | Slow | No | Yes | Yes | Yes |
| Statistical Potential | Fast | No | No | No | Yes |
| Empirical Function | Fast | Yes | No | No | Yes |
| Machine Learning | Fast | Yes | No | No | No |

**Table 3**

Methods for observing fast folding events [Adapted from [118], p. 77]

| Technique | Approximate timescale probed |
|---|---|
| LASER flash photolysis | 100 fs -1 ms |
| Electron-transfer-induced refolding | 1 μs – 1 ms |
| Acoustic relaxation | 1 ns – 1 ms |
| Dielectric relaxation | 1 ns – 1 s |
| LASER T-jump | 1 ns – 100 ms |
| Electrical discharge T-jump | 100 ns – 10 s |
| Mixing | 10 μs - ∞ |
| Pressure-jump | 60 μs – 1 s |
| NMR line broadening | 100 μs – 100 ms |