

# Identifying the importance of amino acids for protein folding from crystal structures

Nikolay V. Dokholyan<sup>1,2\*</sup>, Jose M. Borreguero<sup>3</sup>, Sergey V. Buldyrev<sup>3</sup>,  
Feng Ding<sup>3</sup>, H. Eugene Stanley<sup>3</sup>, and Eugene I. Shakhnovich<sup>1</sup>

<sup>1</sup>*Department of Chemistry and Chemical Biology,  
Harvard University, Cambridge, MA 02138*

<sup>2</sup>*Department of Biochemistry and Biophysics,  
University of North Carolina at Chapel Hill,  
School of Medicine, Chapel Hill, NC 27599 and*

<sup>3</sup>*Center for Polymer Studies, Department of Physics,  
Boston University, Boston, MA 02215*

(Dated: Printed: February 27, 2003)

## Abstract

We review recent advances in determining and characterizing protein folding kinetics from crystal structures using computational techniques. We also describe a new protein model and show that it reproduces the experimentally observed folding thermodynamics and kinetics of SH3 domain, a small protein, that has been experimentally studied in detail. We verify a *nucleation* mechanism as a scenario for the folding of SH3 domain. We identify the transition state ensemble of the SH3 domain and dissect it by quantifying the protein topology using concepts of graph theory.

---

\* Corresponding author. Email: dokh@med.unc.edu

## I. INTRODUCTION

One of the most intriguing questions in biophysics is how protein sequences determine their unique three-dimensional structure. This question, known as the protein folding problem [1–25], is of great importance because understanding protein folding mechanisms is a key to successful manipulation of protein structure and, consequently, function. The ability of manipulating protein function is, in turn, crucial for effective drug discovery.

Understanding the mechanisms of protein folding is also crucial for deciphering the imprints of evolution on protein sequence and structural spaces. For example, some positions along the sequence in a set of structurally similar non-homologous proteins are more conserved in the course of evolution than others [26]. Such conservation can be attributed to evolutionary pressure to preserve amino acids that play a crucial role in: *(i)* protein function, *(ii)* stability, and *(iii)* folding kinetics — the ability of proteins to rapidly reach their native state [27, 28]. Interestingly, function is not conserved among non-homologous proteins that share the same fold, so we can assume that the evolutionary pressure to preserve functionally important amino acids in such a set of proteins is “weaker” than those that are involved in protein stability and folding kinetics. It has been shown [27] that up to 80% of conservation of amino acids in the course of evolution can be explained by pressure to preserve protein stability. Thus, in order to understand the role of evolutionary pressure to preserve rapid folding kinetics, we need to be able to quantify the importance of amino acids for protein folding kinetics.

Due to difficulties and cost of actual experimental studies, it is important to develop rapid tools to identify folding kinetics of a given protein from its crystal structure. The ultimate goal is to be able to predict protein folding kinetics of a given protein from its sequence. However, this goal requires the solution of the protein folding problem (Sec. II), i.e. understanding of how a given amino acid sequence folds into native protein structure. Since protein crystal structures provide invaluable information about amino acid interactions, it is possible to reduce the problem to identifying protein folding kinetics from its structures. Surprisingly, such an approach has already yielded promising and robust results.

Here we present an overview of recent efforts to reconstruct the folding mechanism of proteins using various computational techniques from proteins crystal structures. We also describe our methods that we have recently validated on Src SH3 domain, a 56-amino acid

protein, studied in detail in experiments [16, 22, 29–42] and molecular dynamics simulations [43–47]. We describe a new protein model and show that the thermodynamics of Src SH3 from molecular dynamics simulations is consistent with that experimentally observed. We test the proposed mechanism for protein folding, the *nucleation scenario*, and identify the transition state ensemble of protein conformations, characterized by the maximum of the free energy [7, 13, 15, 48, 49].

## II. HOW DO PROTEINS FOLD?

### A. Levinthal “paradox”

In 1968, Levinthal formulated a simple argument that points out the non-random character of protein folding kinetics [50], which we illustrate with the following example. Consider a 100 amino acid protein and let us estimate the time necessary for such a protein to reach its unique native state by a random search. If each amino acid moves in 6 possible directions (e.g. up, down, left, right, forward, backward), the total number of conformations that a 100-amino acid protein can assume is  $6^{100} \approx 10^{78}$ . It is known that the fastest vibrational mode of a protein is that of its tails and is of the order of magnitude of  $1ps$ , so the time necessary for a 100-amino acid protein to fold is approximately  $10^{66}s$  or  $10^{57}$  years. Many proteins fold in the  $1ms - 1s$  range. Thus, there is a specific mechanism due to which proteins avoid most conformations en route to their native state.

### B. Nucleation scenario

Two-state proteins are characterized by fast folding and the absence of stable intermediates at physiological temperatures. If we follow the folding process for an ensemble of initially unfolded proteins, both the average potential energy and the entropy of the ensemble decrease smoothly to their native state values. The absence of energetic and topological frustrations defines a “good folder” [12, 51]. Various measures have been proposed to determine if a protein sequence qualifies as a two-state folder, either relying on kinetic [52] or thermodynamic [9, 53] properties.

The free energy landscape of the two-state proteins at physiological temperatures is characterized by two deep minima [7, 11, 15, 35, 54–56]. One minimum corresponds to the unique

native state with the lowest potential energy and low conformational entropy, while the second minimum corresponds to a set of unfolded or misfolded conformations with higher values of potential energy and high conformational entropy. At the folding transition temperature  $T_F$ , these minima have equal depths, and both native state and unfolded state coexists in equilibrium with equal probability. The two minima are separated by a free energy barrier. The set of conformations that belong to the top of this barrier, having the maximal values of free energy, is called the *transition state ensemble*.

At equilibrium, the probability of observing a conformation with free energy,  $\Delta G$ , is given by  $p \sim \exp(-\Delta G/k_B T)$ , where  $k_B$  is the Boltzmann constant and  $T$  is the temperature of the system. Since at  $T_F$  free energies of native and unfolded/misfolded ensembles are equal, the probability to exist in each of these states is the same. The probability to find a conformation at the top of the free energy barrier is minimal. Therefore, if we consider any protein conformation at the top of the free energy barrier, such a conformation most likely unfolds or reaches its native state with equal probabilities 1/2. So, the transition state ensemble is characterized by probabilities of the conformations to reach the native state equal to 1/2 [46, 47, 57].

The questions then are: “Which conformations belong to the top of the free energy barrier?” “Are there any specific mechanisms that are responsible for the rapid folding transition?” Numerous folding scenarios have been proposed to answer these questions [6, 13, 58–64]. The mechanism that we advocate in the present paper is called a *nucleation scenario* [10, 11]. According to the nucleation scenario, there is a specific obligatory set of contacts at the transition state ensemble, called a *specific nucleus*, the formation of which determines the future of a conformation at the transition state ensemble. If the specific nucleus is formed, a protein rapidly folds to its native conformation. If the specific nucleus is disrupted in the transition state, the protein rapidly unfolds. Thus, to verify the nucleation scenario we must determine the nucleus and the transition state ensemble of a protein. Next, we describe a protein model that we use in molecular dynamics simulations.

### III. PROTEIN ENGINEERING EXPERIMENTS

A way to test the importance of amino acids in experiments was proposed by Fersht [65, 66]. The method, called *protein engineering* or  $\Phi$ -value analysis, is based on the engineering

a mutant protein with amino acids under consideration replaced by other ones. The value of free energy difference between the wild type and the mutant proteins is measured at the transition state  $\Delta G^\ddagger$ , folded state  $\Delta G^F$ , and unfolded states  $\Delta G^U$  (Fig. 3).  $\Phi$ -values are defined as

$$\Phi = \frac{\Delta G^\ddagger - \Delta G^U}{\Delta G^F - \Delta G^U}. \quad (1)$$

$\Phi$ -values are close to zero for those amino acids whose substitution does not affect the transition states. Thus, at the zeroth approximation, these amino acids are least important for the protein folding kinetics.  $\Phi$ -values are close to unity for those amino acids whose substitution affect the transition states to the same extent as the folded states. Thus, these amino acids are most important for the protein folding kinetics.

#### IV. RECENT DEVELOPMENTS IN THE DETERMINATION OF PROTEIN FOLDING KINETICS

Developments in the last decade in protein purification and structure-refining methods [67–70] have led to publication of high resolution proteins’ crystal structures. This set of data boosted theoretical studies of protein folding beyond the general heteropolymer models [4, 71–75]. Early studies targeting important amino acids for protein dynamics applied the available crystal structure data in two different approaches: structures were used (i) as reference states (decoys) for theoretical predictions [76–83], and (ii) as a source of dynamical information [84–88]. Studies relied in the developed theoretical framework [89] that explained the folding of relatively small proteins as a chemical reaction between two sets of species — folded and denaturated protein states, separated by transition states and by possibly a set of metastable intermediates. Transition states control the rate of the folding reaction, and solving for the portions of the protein that provide structural coherence to these transition states became a major effort in determining the kinetically important amino acids.

Computational power limitation and the inaccuracies in the inter-atomic force-field [90] forced all-atom folding simulations to be performed under extreme conditions favoring denaturation, typically very high temperatures [76–82]. This approach assumes that folding of the protein can be described by running the unfolding simulation backwards in time, and that folding at high temperatures is comparable to folding at room temperatures. These

assumptions are questionable, since folding experimental studies are performed under conditions favoring the native state. Furthermore, the low stability of proteins at physiological conditions —only a few kcal/mol [91], indicates that folding of the protein to its native structure is the result of a delicate balance between enthalpic and entropic terms. This balance is distorted at high temperatures, where folding becomes a rare event and the transition state may change drastically [92].

In simulations, Daggett et al. [77, 79, 81] unfolded target proteins starting from their crystal structures, and monitored the time evolution of a parameter representing the structural integrity of proteins during simulations. Abrupt changes in the parameter pinpointed denaturation of these protein, and analysis of the trajectories revealed disrupted native amino acid interactions. The amino acids involved in these key interactions were identified as kinetically important, and the authors found good correlation to experimental folding results.

The issue of the limited statistical significance of the results [77, 79, 81] due to a small number of unfolding simulations was addressed by Lazaridis et al. [76], who performed a larger series of unfolding simulations starting from conformations slightly different from the initial crystal structure. A wealth of simulations allowed authors to extract the common set of key interactions and identify the important amino acids with higher accuracy. Other attempts to circumvent the poor statistics rested on the discretization of a representative unfolding simulation, followed by long equilibrium simulations of the protein around each of the discretized steps [78, 80]. This method assumes that a protein is at equilibrium at every step in the folding process, but given that at high temperatures folding is a rare event, caution must be taken when interpreting the results.

Recent all-atom simulations were also used to increase the efficiency of protein engineering experiments in a self-consistent experimental+computational approach toward determination of the TSE [82]. This method is most useful for proteins for which only a small fraction of the residues play a key role. Such method may also serve as a refining tool of the protein engineering results.

Protein databases [93] of crystal structures have been widely used as a source of dynamical information with application to folding simulations. In their pioneer study, Wilson et al. [85] computed effective pairwise amino acid contact potentials from the frequencies of spatial proximities between pairs of amino acids obtained in the database of structures. Authors

used these potentials to reproduce with modest success the folding process of a one-atom crambin [94] model on the square lattice. Skolnick and Kolinski [86] developed a statistical potential using two-atom representation of apoplastocyanin [95]. Folding simulations on a finer lattice than that used in previous studies allowed authors to fold a model protein with a root mean square deviation (*rms*) of 6Å with respect to the crystal structure. However, the propensity of the amino acids to adopt a specific crystal structure prevented authors from generalizing the applicability of the model to more than one protein at a time.

Kolinsky et al. [84] extended the original model [86] with a sophisticated potential energy including a variety of energetic and entropic contributions and a hierarchy of finer lattices. Refolding simulations of three different proteins allowed authors to describe folding processes with moderate success. Vieth et al. [88] used a similar model to study the aggregation kinetics of the GCN4 leucine zipper [96] into dimers, trimers and four-mers. Authors identified the amino acids that regulated the aggregation kinetics, in accordance with experiments. A systematic study [97] indicated that simple pairwise statistical potentials are of limited use in refolding simulations, and although statistically derived potentials are gaining in prediction power, their rapidly increasing complexity compromised their efficiency when compared to *ab initio* molecular dynamics simulations.

Since all information necessary to fold a particular protein is precisely encoded in the protein structure, the crystal structure can be used as the sole source of information, with no regard to the protein database. This approach was taken by Dill et al. [87] in their study of the folding mechanisms of crambin and chymotrypsin inhibitor [98]. Dill et al. assigned attractive interactions between all pairs of hydrophobic amino acids that were in geometrical proximity from each other in the crystal structure, neglecting other amino acid interactions. The folding dynamics was implemented through a sequence of folding events in Monte Carlo search. Authors found that one every 4000 simulations ended in the crystal structure and proposed a folding pathway for the two proteins. This technique, although able to find a folding event, cannot reproduce a statistically significant ensemble, since the sequence of folding events is forced in the simulation. Thus, only when the proposed sequence of events coincides with the most probable ones, can the results be representative of the folding of the protein.

Crystal structure-based approaches to identify the important amino acids for protein folding have attracted interests in the past decade [39, 43, 55, 64, 77, 99–101]. These meth-

ods either use the crystal structure as the reference state or use the crystal structure based interaction potentials (Gō potential; see Sec. V). Starting from the crystal structure, temperature induced unfolding [77, 99] in all-atom molecular dynamics simulations with explicit solvent molecules have been applied to study the transition states. However, the limitation of computational ability of traditional molecular dynamics algorithms only enables one to sample over several unfolding trajectories from folded state. Thus, this technique can only capture one or a few transition state conformations instead of a statistically significant ensemble. Moreover, derivation of *folding* transition state ensemble from high temperature unfolding may be problematic in some cases due to possible significant differences between high-temperature free energy landscape and the free energy landscape of a protein at physiological temperatures [92, 102].

Alternative theoretical approaches [39, 55, 64, 100] have been proposed to predict the transition states in protein folding and obtained significant correlations with experimental  $\phi$  values for several proteins. However, each of these models involves drastic assumptions. For example, each amino acids can only adopt two states — native or denatured — and the ability to be in the native state was considered to be independent of other residues. Such an assumption holds for one-dimensional systems, but may be inappropriate for three-dimensional proteins, because the native state of a residue depends on its contacts with its neighbors.

Combined with effective dynamic algorithm, simplified protein models with crystal structure based interaction potential [43, 101, 103] have been applied to study the folding kinetics. The principal difficulty in the kinetics studies is the classification of various protein conformations, i.e. the knowledge of *reaction coordinate* – a parameter that can uniquely identify position of a protein conformation on a folding landscape with respect to the native state. The fraction of native contacts  $Q$  [43, 101] has been proposed as an approximation to the reaction coordinate. However, other authors have argued that the reaction coordinate for folding is not well defined [7, 57, 104], and the principal difficulty in identifying the folding reaction coordinate from crystal structures is in uncovering the relationship between protein folding thermodynamics and kinetics, i.e. how much kinetic information we can obtain about protein folding barriers from *equilibrium* sampling of folding trajectories.

## V. PROTEIN FOLDING KINETICS FROM DISCRETE MOLECULAR DYNAMICS SIMULATIONS

### A. Protein model

The problem of protein modeling in simulations is as complex as the protein folding problem itself. Such complexity often makes unpractical brute-force approaches of all-atom simulations. Lattice models [7, 56, 71, 105–110] became popular due to their ability to reproduce a significant amount of folding events in a reasonable computational time. However, the role of topology in determining the folding nucleus requires study beyond lattice models, which impose unphysical constraints on the protein degrees of freedom. Simplified off-lattice models [43, 49, 111–115] are a compromise between lattice and all-atom models and the first step into modeling the conformational dynamics of proteins.

A simple minimalistic off-lattice protein model is a beads-on-a-string model, representing a chain with maximal flexibility [116]. One drawback of a bead-on-a-string model is that its chain flexibility is higher than that observed in real proteins, so, as the result, the protein model folding kinetics is often altered due the conformational traps that occur in excessively flexible protein models during folding. Stiffer chains allow more cooperative motions of protein chains, drastically reducing the number of collapsed conformations. It is, thus, crucial to introduce an additional set of chain constraints in order to mimic the flexibility of the proteins.

In Ref. [47], we model a protein by beads representing  $C_\alpha$  and  $C_\beta$  (Fig. 1a). There are four types of bonds: (i) covalent bonds between  $C_{\alpha i}$  and  $C_{\beta i}$ , (ii) peptide bonds between  $C_{\alpha i}$  and  $C_{\alpha(i\pm 1)}$ , (iii) effective bonds between  $C_{\beta i}$  and  $C_{\alpha(i\pm 1)}$ , (iv) effective bonds between  $C_{\alpha i}$  and  $C_{\alpha(i\pm 2)}$ . In order to determine the effective bond length, we calculate the average and the standard deviation of distances between carbon pairs of types (iii) and (iv) for  $10^3$  representative globular proteins obtained from the PDB [93]. We find that the average distances are  $4.7\text{\AA}$  and  $6.2\text{\AA}$  for type (iii) and type (iv) bonds, respectively. The ratio  $\sigma$  of the standard deviation to the average for bond types (iii) and (iv) are 0.036 and 0.101, respectively. The standard deviation of bond type (iv) is larger than that of bond type (iii) because it is related to the angle of two consecutive peptide bonds. Thus, the bond lengths of type (iv) fluctuate more than those of type (iii). The effective bonds impose additional

constraints on the protein backbone so that our model closely mimics the stiffness of the protein backbone, and can give rise to cooperative folding thermodynamics.

In our simulation, the four types of bonds are realized by assigning infinitely high potential well barriers [116] (Fig. 1b):

$$V_{ij}^{\text{bond}} = \begin{cases} 0, & D_{ij}(1 - \sigma) < |r_i - r_j| < D_{ij}(1 + \sigma) \\ +\infty, & \text{otherwise} \end{cases}, \quad (2)$$

where  $D_{ij}$  is the distance between atoms  $i$  and  $j$  in the native state,  $\sigma = 0.0075$  for a bond of type  $(i)$ ,  $\sigma = 0.02$  for a bond of type  $(ii)$ ,  $\sigma = 0.036$  for a bond of type  $(iii)$  and  $\sigma = 0.101$  for a bond of type  $(iv)$ . The covalent and peptide bonds are given a smaller width and the effective bonds are given a larger width to mimic the protein flexibility. Other models tailored for molecular dynamics include the use of continuous potentials for bond and dihedral angles [43, 114, 117] and for distances [118]. However, the use of discrete potential of interactions presents a computational simplification over continuous potentials that require calculations every discrete time step.

We use a modified Gō model similar to one described in [116], in which interactions are determined by the native structure of proteins. In our model, only  $C_\beta$  atoms that are not nearest neighbours along the chain interact with each other. The cutoff distance between  $C_\beta$  atoms is chosen to be 7.5Å. The Gō model has been widely applied to study various aspects of protein folding thermodynamics and kinetics [24, 46, 47, 49, 64, 119, 120].

Despite the drawback of the Gō model, associated with the prerequisite knowledge of the native structure, it has important advantages. It is the simplest model that satisfies the principal thermodynamic and kinetic requirements for a protein-like model:  $(i)$  the unique and stable native state,  $(ii)$  a cooperative folding transition resembling a first-order phase transition. Importantly, protein sequences with amino acids represented by only two or three types showed at  $T_F$  a fast decrease of the potential energy, followed by a slow decrease of the energy until proteins reached their native state. The corresponding folding scenario is a coil-to-globule collapse, followed by a slow search of the native structure through metastable intermediates [113, 117, 121]. Similarly, the addition of non-specific interactions to the Gō model resulted in analogous trapping [114].  $(iii)$  The Gō model is derived from the native topology, which according to protein engineering experiments [16, 17, 122, 123] is determinant in the resulting structure of the transition state. Furthermore, in a recent study

with an all-atom energy function, Paci et al. determined that native interactions account for 85% of the energy of the transition state ensemble of the two-state folder AcP [124].

The use of the  $G\bar{o}$  model is based (implicitly or explicitly) on the assumption that topology of the native structure is more important in determining folding mechanism than energetics of actual sequences that fold into it. Apparently, a conclusive proof of such assumption can be obtained either in simulations which do not use the  $G\bar{o}$  model, or in experiments that compare folding pathways of analogs — proteins with non-homologous sequences that fold into similar conformations. The dominating role of topology in defining folding mechanism was first found in simulations in 1994 when Abkevich and coauthors [7] observed that various nonhomologous sequences designed to fold to the same lattice structure featured the same folding nucleus. This finding was further corroborated in [125] where evolution-like selection of fast-folding sequences generated many families of sequences (akin to superfamilies in real proteins) that all have the same nucleus positions, stabilized despite the fact that actual aminoacid types that delivered such stabilization varied from family to family. Similar behavior was observed in structural and sequence alignment analysis of real proteins [26, 125, 126], where extra conservation was detected in positions corresponding to common folding nucleus for proteins representing that fold. Experimentally common folding nucleus was found in  $\alpha/\beta$  plait proteins that have no sequence homology [122, 127]. Other works provided support of the important role of protein topology to its folding kinetics [43, 122, 128–130].

## B. Discrete Molecular Dynamics algorithm

Due to the computational burden of traditional molecular dynamics [131], simplified simulation methods are needed to study protein folding. Our program employs the discrete molecular dynamics algorithm, which recently received strong attention due to its rapid performance [132, 133] in simulating polymer fluids [132], single homopolymers [133, 134], proteins [116, 119, 135], and protein aggregates [136, 137]. The detailed description of the algorithm can be found in [138–141].

To control the temperature of the protein we introduce  $\sim 10^3$  particles, which do not interact with the protein or with each other in any way but via elastic collisions, serving as a heat bath. Thus, by changing the kinetic energy of those “ghost” particles we are able to control the temperature of the environment. The “ghost” particles are hard spheres of the

same radii as chain residues and have unit mass. Throughout the paper the temperature is given in units of  $\epsilon/k_B$ . The time unit (tu) is estimated from the shortest time between two consecutive collisions between any two particles in the system.

### C. Folding thermodynamics

To test whether our models faithfully reproduce the experimentally-observed [16, 34, 35, 142] thermodynamic and kinetic properties of SH3 domain, Ding et al. [47] and Borreguero et al. [46] performed the discrete molecular dynamics simulations of the model SH3 domain at various temperatures. At each temperature we calculate the potential energy  $E$ , the radius of gyration  $R_g$  [143], the root-mean-square deviation from the native state  $RMSD$  [144], and the specific heat  $C_v(T)$ . The radius of gyration is a measure of a protein size,  $RMSD$  measures the similarity between a given conformations and the native state, and the specific heat measures the fluctuations of the potential energy of the protein at a given temperature.

At low temperatures, the average potential energy  $\langle E \rangle$  increases slowly with temperature, and the  $RMSD$  remains below  $3\text{\AA}$ . Near the transition temperature  $T_f$ , the quantities  $E$ ,  $R_g$ , and  $RMSD$  fluctuate between values characterizing two states, folded and unfolded, yielding a bimodal distribution of the potential energy. Potential energy fluctuations at  $T_f$  give rise to a sharp peak in  $C_v(T)$  (see e.g. Fig. 7 of ref. [116]), which is characteristic of a first order phase transition for a finite system. Our findings are consistent with the two-state folding thermodynamics, experimentally observed for C-Src SH3 domain [16, 34, 35, 142].

### D. Protein folding kinetics

#### 1. Identifying the folding nucleus

A method to identify the protein folding nucleus from equilibrium trajectories was proposed in Ref. [116] and later used on SH3 domain proteins [46, 47]. The idea is to study ensembles of conformations that have a specific history and future. For example, conformations that originate in the unfolded state, reach a putative transition state, and later unfold, must differ from the conformations that originate in the folded state, reach a putative transition region and later fold. Both sets of conformations, which we denote by UU

and FF, are characterized by the same potential energy and similar overall structural characteristics. Nevertheless, there is a crucial kinetic difference between them. According to nucleation scenario, UU conformations lack the folding nucleus. The nucleus is not created at the transition region, which leads to the protein unfolding. FF conformations have the nucleus intact at the transition state, so that the protein does not unfold. Thus, in order to determine the nucleus, we propose to compare the average frequencies of contacts between amino acids in UU and FF ensembles of conformations. Amino acid contacts that have the largest frequency difference form the folding nucleus.

We test this method to identify the nucleus of a computationally designed protein [49] and later, in Ref. [47], to determine the folding nucleus of Src SH3 domain protein. For SH3 domain we find that the crucial contact that is formed at the top of the free energy barrier is between two loops — the distal hairpin and the divergent turn, namely L24-G54. The observation of this contact is statistically significant: the probability of observing L24-G54 contact in our molecular dynamic simulations *by chance* is about 0.04, although L24-G54 is the most persistent contact in FF-UU ensemble. The formation of this contact clips the distal hairpin and the RT-loop together, drastically reducing protein entropy.

To additionally test the role of contact L24-G54 in SH3, we “covalently” constrain this contact in our molecular dynamics simulations. If L24-G54 constitutes the folding nucleus, then by constraining it we do not allow the folding nucleus to be disrupted, and, thus, the protein should rarely unfold in equilibrium simulations. After cross-linking L24-G54, we observe that SH3 domain exists predominantly in the folded state. In fact, the histogram of the potential energy states, being bimodal at  $T_F$  for unconstrained protein, becomes unimodal, with a maximum corresponding to the energy of the native conformation. Thus, cross-linking of L24-G54 strongly biases conformations to the native state.

For control, we test if constraining any other contact leads to a similar bias of conformations to the native state. We cross-link T9-S64, the N- and C-termini of Src SH3. T9-S64 is the longest range contact along the protein chain, and, in a case of a homopolymer it reduces the entropy of conformational space the most [145]. We find that fixation of T9-S64 does not affect the distribution of energy states, indicating that formation of an arbitrary contact is not a sufficient condition to bias the protein conformation towards its native state.

## 2. Identifying the Transition State Ensemble

Next, we identify the transition state ensemble of SH3 protein — the set of all conformations that belong to the top of the free energy barrier. We test if selected conformations belong to the transition state ensemble by computing its probability to fold,  $p_{\text{FOLD}}$  [57]. To determine  $p_{\text{FOLD}}$  for a given conformation in molecular dynamics simulations, we randomize the velocities of the particles and simulate the protein for a fixed interval of time, long enough to observe a folding transition in equilibrium simulations at  $T_F$ . We then determine  $p_{\text{FOLD}}$  by computing the ratio of number of successful folding events versus total number of trials. As we mention above, transition state ensemble conformations are characterized by  $p_{\text{FOLD}}$  values close to  $1/2$ .

We study three types of conformations: (a) UU, (b) FF, and (c) UF. The later is a set of conformations that originate in the unfolded state, crosses the putative transition barrier, and reaches the folded state. We choose the putative transition conformations as those having an energy higher than that of the native state, lower than the average energy of unfolded states, and having the lowest probability at  $T_F$  (Fig. 2). In UU conformations, the nucleus is not present, and since there is little chance that it will be created after randomization of velocities, we expect  $p_{\text{FOLD}}$  to be close to zero. In FF conformations, the nucleus is present and since there are little chances that it will be disrupted, we expect  $p_{\text{FOLD}}$  to be close to unity. In UF conformations the nucleus is present with some probability, thus, if we select UF conformations so that the nucleus is formed with the probability  $1/2$ , we expect  $p_{\text{FOLD}}$  to be close to  $1/2$ .

In Refs. [46, 47] we show that, in fact,  $p_{\text{FOLD}}$  is close to zero for the ensemble of UU conformations.  $p_{\text{FOLD}}$  is approximately unity for the ensemble of FF conformations. Only for the ensemble of UF conformations we find that  $p_{\text{FOLD}}$  is close to  $1/2$ . Thus, the set of UF conformations represent the transition state ensemble.

It is important, that even though we perform thermodynamic simulations, we study the protein folding kinetics because we select UU, FF and UF conformations based on their past and future states. It is due to kinetic selection of the UU, FF, and UF conformations we observe difference in  $p_{\text{FOLD}}$  values, even though their energetic (potential energy) and structural (RMSD,  $R_g$ ) characteristics are close to each other.

### 3. “Virtual screening” method

We use a technique similar to experimental  $\Phi$ -value analysis to predict the TSE via computer simulations. We assume that the mutation does not give rise to significant variation of the three-dimensional structures of folded and transition state ensembles, the same assumption that is made in protein engineering experiments. In our simulations, the free energy shifts due to mutation can be computed separately in the unfolded, transition, and folded state ensembles:

$$\Delta G_x = -kT \ln \langle \exp(-\Delta E/kT) \rangle_x. \quad (3)$$

Here  $x$  denotes a state ensemble (folded, F, unfolded, U, and transition, ‡),  $\Delta E$  is the change of potential energy due to the mutation, and the average  $\langle \dots \rangle_x$  is taken over all conformations of unfolded, transition, and folded state ensembles. We compute [43]

$$\Phi = \frac{\ln \langle \exp(-\Delta E/kT) \rangle_{\ddagger} - \ln \langle \exp(-\Delta E/kT) \rangle_U}{\ln \langle \exp(-\Delta E/kT) \rangle_F - \ln \langle \exp(-\Delta E/kT) \rangle_U}. \quad (4)$$

The  $\Phi$ -values in our analysis are determined using the free energy relationship of Eq.(3) that takes into account both energetic and entropic contributions, but assumes that mutations do not change the TSE. Interestingly, if one adopts a simplified definition of  $\Phi$ -value used in recent work [146] as proportional to the number of contacts a residue makes in the TSE, the correlation coefficient between theoretical and experimental  $\Phi$ -values is reduced to 0.27 from approximately 0.6 (Sec. VD 4). An approximation to the  $\Phi$ -value, the *difference* between the average number of contacts residues form in the TSE and in unfolded states,  $\Phi \approx (\langle N_i \rangle_{\ddagger} - \langle N_i \rangle_U) / (\langle N_i \rangle_F - \langle N_i \rangle_U)$ , provides a better correlation coefficient between predicted and experimentally observed  $\Phi$ -values (0.48) than does the approximation of Ref. [146]. The reason that a thermodynamic definition of the  $\Phi$ -value yields better agreement with experiments can be inferred from a  $\Delta G$  plot [47], which shows that  $\Delta G^F - \Delta G^U$  for most of the amino acids is *not* negligible. Indeed, there are several amino acids that make persistent short-range contacts in the unfolded states.

### 4. Comparing simulations to experiments

In Ref. [47],  $\Phi$ -values are computed using the *virtual screening method* and the comparison with experimental  $\Phi$ -values [16, 34] for Src SH3 protein was statistically significant — the

linear regression coefficient is approximately 0.6. By comparing the number of contacts that an amino acid makes in the TSE with that number in the unfolded state, those amino acids that are most important for the formation of the transition state ensemble are selected: L24, F26, L32, V35, W43, A45, A54, Y55 and I56. In general, the majority of the residues from that list have high experimental  $\Phi$ -values; remarkably, residue A45 which has the highest number of contacts in the transition state ensemble with respect to the unfolded states has the highest experimental  $\Phi$ -value — 1.2. Notable exceptions are L24, W43 and G54, which have  $\Phi$ -values that are either small or negative, as in the case of G54.

For residue G54, mutation destabilizes the protein while accelerating folding, strongly suggesting that it participates in the transition state ensemble [147]. Additional evidence supporting the important roles of L24 and G54 for the transition state of SH3 comes from the evolutionary observation that these amino acids are conserved in a family of homologous SH3 domain proteins [46, 148].

## VI. ROLE OF PROTEIN TOPOLOGY

*En route* to the native state, at the transition states a protein loses its entropy by forming a specific nucleus. Entropically and energetically pre- and post-transition states — conformations with  $p_{\text{FOLD}}$  approximately zero and unity correspondingly — are indistinguishable. In fact, in Ref. [120], pre- and post-transition sets of conformations were selected for SH3 and CI2 proteins. Both pre- and post-transition states had similar structural and energetic properties. The question then is: “What distinguishes pre- and post-transition states?”

To answer this question, we hypothesize that the actual topological properties of pre- and post-transition conformations are different. To test this hypothesis, we construct a protein graph, nodes of which represent amino acids, and edges represent pairs of amino acids that are within the contact range from each other. For SH3, we define the maximum distance between  $C_\beta$  atoms at which a contact exists at  $7.5\text{\AA}$  [149].

A simple measure of topological properties of the graph is the average minimal path along the edges between any two nodes of the graph,  $L$ , used recently in Ref. [150] and later used for discriminating pre- and post-transition states of SH3 and CI2 proteins:

$$L = \frac{1}{N(N-1)} \sum_{i>j}^N \ell_{ij}, \tag{5}$$

where  $N$  is the number of amino acids,  $\ell_{ij}$  is the minimal path between nodes  $i$  and  $j$ .  $L$ -values characterize the “tightness” of the network by computing the average separation of elements from each other.

For both SH3 and CI2 proteins,  $L$  was observed to be significantly different, supporting our hypothesis that the protein conformation topology plays an important role in protein folding kinetics. Additional evidence of the importance of topology in protein folding was shown in Ref. [150], where using other determinants of protein graph topology, the most important amino acids for the protein folding kinetics were identified for several proteins: AcP, human procarboxypeptidase A2, tyrosine-protein kinase SRC,  $\alpha$ -spectrin SH3 domain, CI2, and protein L. Using Monte-Carlo simulations of hydrophobic protein model, Treptow et al. [151] also suggested the role of protein topology in folding kinetics.

## VII. CONCLUSION

We describe recent advances in determining and characterizing protein folding kinetics from crystal structures using a variety of analytical and computational tools. We describe a protein model for off-lattice molecular dynamic simulations that faithfully reproduces many aspects of SH3 folding thermodynamics and kinetics. Using Molecular Dynamics simulations, we verify the *nucleation scenario* for the SH3 protein family by comparing the fluctuations originating at the native and unfolded states. We find an important role of L24-G54 contact for the folding kinetics of SH3 proteins. A possible test of kinetic importance of the L24-G54 contact may come from cross-linking this contact and understanding if cross-linking stabilizes the native state of the Src SH3.

We identify the transition state ensemble for Src SH3 protein, and find that it is consistent with experimental observations. We dissect the transition state ensemble by studying wiring properties of protein graphs. The structural properties of protein graphs is related to protein topology and, thus, may explain the kinetics of the folding process.

## VIII. ACKNOWLEDGMENTS

We greatly acknowledge E. Deeds and A. F. P. de Araújo for their critical reading of the manuscript. This work is supported by NSF and Petroleum Research Fund (to HES), and

NIH (GM52126 to ES). NVD is supported by NIH NRSA Grant (GM20251-01).

---

- [1] Anifsen, C. B., Principles that govern the folding of the protein chains, *Science* **181**, 223–230 (1973)
- [2] Taketomi, H., Ueda, Y. & Gō, N., Studies on protein folding, unfolding and fluctuations by computer simulations, *Int. J. Peptide Protein Res.* **7**, 445 (1975)
- [3] Gō, N., Theoretical studies of protein folding, *Ann. Rev. Biophys. Bioeng.* **12**, 183–210 (1983)
- [4] Bryngelson, J. D. & Wolynes, P. G., Intermediates and barrier crossing in a random energy model (with applications to protein folding), *J. Phys. Chem.* **93**, 6902–6915 (1989)
- [5] Karplus, M. & Shakhnovich, E. I. Protein folding: theoretical studies of thermodynamics and dynamics. in Creighton, T., ed., *Protein Folding*. W. H. Freeman and Company, New York 1994.
- [6] Ptitsyn, O. B., Kinetic and equilibrium intermediates in protein folding, *Protein Eng.* **7**, 593–596 (1994)
- [7] Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I., Specific nucleus as the transition state for protein folding: evidence from the lattice model, *Biochemistry* **33**, 10026–10036 (1994)
- [8] Shakhnovich, E. I., Abkevich, V. I. & Ptitsyn, O., Conserved residues and the mechanism of protein folding, *Nature* **379**, 96–98 (1996)
- [9] Klimov, D. K. & Thirumalai, D., Criterion that determines the foldability of proteins, *Phys. Rev. Lett.* **76**, 4070–4073 (1996)
- [10] Fersht, A. R., Nucleation mechanisms in protein folding, *Curr. Opinion Struc. Biol.* **7**, 3–9 (1997)
- [11] Shakhnovich, E. I., Theoretical studies of protein-folding thermodynamics and kinetics, *Curr. Opinion Struc. Biol.* **7**, 29–40 (1997)
- [12] Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G., Theory of protein folding: The energy landscape perspective, *Ann. Rev. Phys. Chem.* **48**, 545–600 (1997)
- [13] Shakhnovich, E. I., Protein design: a perspective from simple tractable models, *Folding & Design* **3**, R45–R58 (1998)
- [14] Micheletti, C., Banavar, J. R., Maritan, A. & Seno, F., Protein structures and optimal folding emerging from a geometrical variational principle, *Phys. Rev. Lett.* **82**, 3372–3375 (1998)

- [15] Pande, V. S., Grosberg, A. Yu., Rokhsar, D. S. & Tanaka, T., Pathways for protein folding: is a new view needed?, *Curr. Opinion Struct. Biol.* **8**, 68–79 (1998)
- [16] Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D., Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain, *Nature Struct. Biol.* **5**, 714–720 (1998)
- [17] Martinez, J. C., Pissabarro, M. T. & Serrano, L., Obligatory steps in protein folding and the conformational diversity of the transition state, *Nature Struct. Biol.* **5**, 721–729 (1998)
- [18] Chan, H. S. & Dill, K. A., Protein folding in the landscape perspective: Chevron plots and non-Arrhenius kinetics, *Proteins: Struct. Func. Genet.* **30**, 2–33 (1998)
- [19] de Araújo, A. F. P., Folding protein models with a simple hydrophobic energy function: The fundamental importance of monomer inside/outside segregation, *Proc. Natl. Acad. Sci. U. S. A.* **96**, 12482–12487 (1999)
- [20] Dinner, A. R. & Karplus, M., The thermodynamics and kinetics of protein folding: A lattice model analysis of multiple pathways with intermediates, *J. Chem. Phys.* **37**, 7976–7994 (1999)
- [21] Bursulaya, B. D. & Brooks, C. L., Folding free energy surface of a three-stranded beta-sheet protein, *J. Am. Chem. Soc.* **121**, 9947–9951 (1999)
- [22] Nölting, B. & Andert, K., Mechanism of protein folding, *Proteins: Struct. Func. Genet.* **41**, 288–298 (2000)
- [23] Ozkan, S. B., Bahar, I. & Dill, K. A., Transition states and the meaning of  $\phi$ -values in protein folding kinetics, *Nature Struct. Biol.* **8**, 765–769 (2001)
- [24] Dokholyan, N. V., The protein folding problem, *Recent Res. Develop. Stat. Phys.* **1**, 77–84 (2001)
- [25] Finkelstein, A. V. & Ptitsyn, O. B., eds., *Protein physics: A course of lectures*, Academic Press, Boston, 2002
- [26] Mirny, L. A. & Shakhnovich, E. I., Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function, *J. Mol. Biol.* **291**, 177–196 (1999)
- [27] Dokholyan, N. V. & Shakhnovich, E. I., Understanding hierarchical protein evolution from first principles, *J. Mol. Biol.* **312**, 289–307 (2001)
- [28] Dokholyan, N. V., Mirny, L. A. & Shakhnovich, E. I., Understanding conserved amino acids

- in proteins, *Physica A* **314**, 600–606 (2002)
- [29] Combs, A. P., Kapoor, T. M., Feng, S. & Chen, J. K., Protein structure–based combinatorial chemistry: discovery of non–peptide binding elements to Src SH3 domain, *J. Am. Chem. Soc.* **118**, 287–288 (1996)
- [30] Feng, S., Chen, J. K., Yu, H., Simons, J. A. & Schreiber, S. L., Two binding orientations for peptides to the Src SH3 domain: development of a general model for SH3–ligand interactions, *Science* **266**, 1241–1247 (1994)
- [31] Feng, S., Kasahara, C., Rickles, R. J. & Schreiber, S. L., Specific interactions outside the proline–rich core of two classes of Src homology 3 ligands, *Proc. Natl. Acad. Sci. U. S. A.* **92**, 12408–12415 (1995)
- [32] Yu, H., Rosen, M. K., Shin, T. B., Seidel-Dugan, C. & Brugde, J. S., Solution structure of the SH3 domain of Src and identification of its ligand-binding site, *Science* **258**, 1665–1668 (1992)
- [33] Grantcharova, V. P. & Baker, D., Folding dynamics of the Src SH3 domain, *Biochemistry* **36**, 15685–15692 (1997)
- [34] Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I. & Baker, D., Experiment and theory highlight role of native state topology in SH3 folding, *Nature Struct. Biol.* **6**, 1016–1024 (1999)
- [35] Viguera, A. R., Martinez, J. C., Filimonov, V. V., Mateo, P. L. & Serrano, L., Thermodynamic and kinetic-analysis of the SH3 domain of Spectrin shows a 2-state folding transition, *Biochemistry* **33**, 10925–10933 (1994)
- [36] Viguera, A. R., Serrano, L. & Wilmanns, M., Different folding transition states may result in the same native structure, *Nature Struct. Biol.* **10**, 874–880 (1996)
- [37] Martinez, J. C., Viguera, A. R., Berisio, R., Wilmanns, M., Mateo, P. L., Filimonov, V. V. & Serrano, L., Thermodynamic analysis of alpha-spectrin SH3 and two of its circular permutants with different loop lengths: Discerning the reasons for rapid folding in proteins, *Biochemistry* **38**, 549–559 (1999)
- [38] Knapp, S., Mattson, P. T., Christova, P., Berndt, K. D., Karshikoff, A., Vihinen, M., Smith, C. I. & Ladenstein, R., Thermal unfolding of small proteins with SH3 domain folding pattern, *Proteins: Struc. Func. Genet.* **23**, 309–319 (1998)
- [39] Guerois, R. & Serrano, L., The SH3-fold family: Experimental evidence and prediction of

- variations in the folding pathways, *J. Mol. Biol.* **304**, 967–982 (2000)
- [40] Mok, Y.-K., Elisseeva, E. L., Davidson, A. R. & Forman-Kay, J. D., Dramatic stabilization of an SH3 domain by a single substitution: Roles of the folded and unfolded states, *J. Mol. Biol.* **307**, 913–928 (2001)
- [41] Northey, J. G. B., Nardo, A. A. Di & Davidson, A. R., Hydrophobic core packing in the SH3 domain folding transition state, *Nature Struct. Biol.* **9**, 126–130 (2002)
- [42] Northey, J. G. B., Maxwell, K. L. & Davidson, A. R., Protein folding kinetics beyond the  $\Phi$  value: Using multiple amino acid substitutions to investigate the structure of the SH3 domain folding transition state, *J. Mol. Biol.* **320**, 389–402 (2002)
- [43] Clementi, C., Nymeyer, H. & Onuchic, J. N., Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins, *J. Mol. Biol.* **278**, 937–953 (2000)
- [44] Gsponer, J. & Catfilsch, A., Role of native topology investigated by multiple unfolding simulations of four SH3 domains, *J. Mol. Biol.* **309**, 285–298 (2001)
- [45] Clementi, C., Jennings, P. A. & Onuchic, J. N., Prediction of folding mechanism for circular-permuted proteins, *J. Mol. Biol.* **311**, 879–890 (2001)
- [46] Borreguero, J. M., Dokholyan, N. V., Buldyrev, S., Stanley, H. E. & Shakhnovich, E. I., Thermodynamics and folding kinetics analysis of the SH3 domain from Discrete Molecular Dynamics, *J. Mol. Biol.* **318**, 863–876 (2002)
- [47] Ding, F., Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I., Direct molecular dynamics observation of protein folding transition state ensemble, *Biophys. J.* **83**, 3525–3532 (2002)
- [48] Gutin, A. M., Abkevich, V. I. & Shakhnovich, E. I., A protein engineering analysis of the transition state for protein folding: simulation in the lattice model, *Folding & Design* **3**, 183–194 (1998)
- [49] Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I., Identifying the protein folding nucleus using molecular dynamics, *J. Mol. Biol.* **296**, 1183–1188 (2000)
- [50] Levinthal, C., Are there pathways for protein folding?, *J. Chim. Phys.* **65**, 44 (1968)
- [51] Onuchi, J. N., Nymeyer, H., Garcia, A. E., Chahine, J. & Socci, N. D., Insights into folding mechanisms and scenarios, *Adv. Prot. Chem.* **53**, 87–152 (2000)

- [52] Bryngelson, J. D., Onuchic, J. N., Socci, N. D. & Wolynes, P. G., Funnels, pathways, and the energy landscape of protein folding - A synthesis, *Proteins: Struct. Func. Genet.* **21**, 167–195 (1995)
- [53] Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I., Improved design of stable and fast-folding model proteins, *Folding & Design* **1**, 221–230 (1996)
- [54] Jackson, S. E., Elmasry, N. & Fersht, A. R., Structure of the hydrophobic core in the transition-state for folding of Chymotrypsin inhibitor-2 — a critical test of the protein engineering method of analysis, *Biochemistry* **32**, 11270–11278 (1993)
- [55] Galzitskaya, O. V. & Finkelstein, A. V., A theoretical search for folding/unfolding nuclei in three-dimensional protein structures, *Proc. Natl. Acad. Sci. U. S. A.* **96**, 11299–11304 (1999)
- [56] Doye, J. P. K. & Wales, D. J., On potential energy surfaces and relaxation to the global minimum, *J. Chem. Phys.* **105**, 8428–8445 (1996)
- [57] Du, R., Pande, V. S., Grosberg, A. Yu., Tanaka, T. & Shakhnovich, E. I., On the transition coordinate for protein folding, *J. Chem. Phys.* **108**, 334–350 (1998)
- [58] Ptitsyn, O. B., Stage mechanism of the self-organization of protein molecules, *Dokl. Acad. Nauk.* **210**, 1213–1215 (1973)
- [59] Kim, P. S. & Baldwin, R. L., Intermediates in the folding reactions of small proteins, *Ann. Rev. Biochem.* **59**, 631–660 (1994)
- [60] Karplus, M. & Weaver, D. L., Protein folding dynamics — the diffusion-collision model and experimental data, *Protein Science* **3**, 650–668 (1994)
- [61] Wetlaufer, D. B., Nucleation, rapid folding, and globular interchain regions in proteins, *Proc. Natl. Acad. Sci. U. S. A.* **70**, 691–701 (1973)
- [62] Wetlaufer, D. B., Nucleation in protein folding — confusion of structure and process, *Trends Biochem. Sci.* **15**, 414–415 (1990)
- [63] Ptitsyn, O. B., How molten is the molten globule?, *Nature Struct. Biol.* **3**, 488–490 (1996)
- [64] Alm, E. & Baker, D., Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures, *Proc. Natl. Acad. Sci. U. S. A.* **96**, 11305–11310 (1999)
- [65] Matouschek, A., Kelis, J. T. Jr., Serrano, L. & Fersht, A. R., Mapping the transient folding intermediates characterized by protein engineering, *Nature* **342**, 122–126 (1989)
- [66] Matouschek, A., Kelis, J. T. Jr., Serrano, L., Bycroft, M. & Fersht, A. R., Transient folding intermediates characterized by protein engineering, *Nature* **346**, 440–445 (1990)

- [67] Ducruix, A. & Geigé, R., *Crystallization of nucleic acids and proteins, a practical approach*, Oxford University Press, Oxford, 1991
- [68] Scopes, R. K., *Protein Purification: Principles and Practice*, Springer Verlag, New York, 1994
- [69] Drenth, J., *Principles of protein x-ray crystallography*, Springer-Verlag, New York, 1994
- [70] Rosenberg, I. A., *Protein analysis and purification: benchtop techniques*, Springer-Verlag, New York, 1996
- [71] Camacho, C. J. & Thirumalai, D., Kinetics and thermodynamics of folding in model proteins, *Proc. Natl. Acad. Sci. U. S. A.* **90**, 6369–6372 (1993)
- [72] Gutin, A. M. & Shakhnovich, E. I., Ground state of random copolymers and the discrete random energy model, *J. Chem. Phys.* **98**, 8174–8177 (1993)
- [73] Pande, V. S., Grosberg, A. Yu. & Tanaka, T., Non-Randomness in Protein Sequences: Evidence for a Physically Driven Stage of Evolution?, *Proc. Natl. Acad. Sci. U. S. A.* **91**, 12972–12975 (1994)
- [74] Chan, H. S. & Dill, K. A., Polymer principles in protein-structure and stability, *Ann. Rev. Biochem.* **20**, 447–490 (1991)
- [75] Bryngelson, J. D., When is a potential accurate enough for structure prediction — Theory and application to a random heteropolymer model of protein-folding, *J. Chem. Phys.* **100**, 6038–6045 (1994)
- [76] Lazaridis, T. & Karplus, M., “New view” of protein folding reconciled with the old through multiple unfolding simulations, *Science* **278**, 1928–1931 (1997)
- [77] Li, A. J. & Daggett, V., Characterization of the transition-state of protein unfolding by use of molecular-dynamics – Chymotrypsin inhibitor-2, *Proc. Natl. Acad. Sci. U. S. A.* **91**, 10430–10434 (1994)
- [78] Boczko, E. M. & Brooks, C. L., First-principles calculation of the folding free-energy of a 3-helix bundle protein, *Science* **269**, 393–396 (1995)
- [79] Daggett, V., Li, A. J., Itzhaki, L. S., Otzen, D. E. & Fersht, A. R., Structure of the transition state for folding of a protein derived from experiment and simulation, *J. Mol. Biol.* **257**, 430–440 (1996)
- [80] Sheinerman, F. B. & III, C. L. Brooks, Molecular picture of folding of a small alpha/beta protein, *Proc. Natl. Acad. Sci. U. S. A.* **95**, 1562–1567 (1998)

- [81] Kazmirski, S. L. & Daggett, V., Non-native interactions in protein folding intermediates: Molecular dynamics simulations of hen lysozyme, *J. Mol. Biol.* **284**, 793–806 (1998)
- [82] Ladurner, A. G., Itzhaki, L. S., Daggett, V. & Fersht, A. R., Synergy between simulation and experiment in describing the energy landscape of protein folding, *Proc. Natl. Acad. Sci. U. S. A.* **95**, 8473–8478 (1998)
- [83] Marti-Renom, N. A., Stote, R. H., Querol, E., Aviles, F. X. & Karplus, M., Refolding of potato carboxypeptidase inhibitor by molecular dynamics simulations with disulfide bond constraints, *J. Mol. Biol.* **284**, 145–172 (1998)
- [84] Kolinski, A. & Skolnick, J., Monte-Carlo simulation of protein folding. Lattice model and interaction scheme., *Proteins: Struct. Func. Genet.* **18**, 338–352 (1994)
- [85] Wilson, C. & Doniach, S., A computer model to dynamically simulate protein folding: Studies with Crambin, *Proteins* **6**, 193–209 (1989)
- [86] Skolnick, J. & Kolinski, A., Simulation of the folding of a globular protein, *Science* **250**, 1121–1125 (1990)
- [87] Dill, K. A., Fiebig, K. M. & Chan, H. S., Cooperativity in protein-folding kinetics, *Proc. Natl. Acad. Sci. U. S. A.* **90**, 1942–1946 (1993)
- [88] Vieth, M., Kolinski, A., Brooks, C. L. & Skolnick, J., Prediction of quaternary structure of coiled coils — application to mutants of the gcn4 leucine-zipper, *J. Mol. Biol.* **251**, 448–467 (1995)
- [89] Shakhnovich, E. I. & Gutin, A. M., Formation of unique structure in polypeptide chains: Theoretical investigation with the aid of a replica approach, *Biophys. Chem.* **34**, 187–199 (1989)
- [90] Wang, W., Donini, O., Reyes, C. M. & Kollman, P. A., Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions, *Ann. Rev. Biophys. Biophys. Struct.* **30**, 211–243 (2001)
- [91] Creighton, T., *Proteins: structures and molecular properties*, W. H. Freeman and Co., New York, ii edition, 1993
- [92] Finkelstein, A. V., Can protein unfolding simulate protein folding?, *Protein Eng.* **10**, 843–845 (1997)
- [93] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov,

- I. N. & Bourne, P. E., The Protein Data Bank, *Nucl. Acid Res.* **28**, 235–242 (2000)
- [94] Teeter, M. M., Roe, S. M. & Heo, N. H., Atomic resolution (0.83Å) crystal structure of the hydrophobic protein Crambin at 130 K, *J. Mol. Biol.* **230**, 292–311 (1993)
- [95] Garrett, T. P., Clingeffer, D. J., Guss, J. M., Rogers, S. J. & Freeman, H. C., The crystal structure of poplar apoplastocyanin at 1.8Å resolution. The geometry of the copper-binding site is created by the polypeptide., *J. Biol. Chem.* **259**, 2822–2834 (1984)
- [96] O’Shea, E. K., Klemm, J. D., Kim, P. S. & Alber, T., X-ray structure of the GCN4 leucine zipper, a two-stranded, parallel coiled coil, *Science* **254**, 539–551 (1991)
- [97] Thomas, D., Casari, G. & Sander, C., The prediction of protein contacts from multiple sequence alignments, *Protein Eng.* **9**, 941–948 (1996)
- [98] McPhalen, C. A. & James, M. N., Crystal and molecular structure of the serine proteinase inhibitor CI2 from barley seeds, *Biochemistry* **26**, 261–278 (1987)
- [99] Li, H., Tang, C. & Wingreen, N. S., Are protein folds atypical?, *Proc. Natl. Acad. Sci. U. S. A.* **95**, 4987–4990 (1998)
- [100] Munoz, V. & Eaton, W. A., A simple model for calculating the kinetics of protein folding from three-dimensional structures, *Proc. Natl. Acad. Sci. U. S. A.* **96**, 11311–11316 (1999)
- [101] Nymeyer, H., Socci, N. D. & Onuchic, J. N., Landscape approaches for determining the ensemble of folding transition states: Success and failure hinge on the degree of frustration, *Proc. Natl. Acad. Sci. U. S. A.* **97**, 634–639 (2000)
- [102] Dinner, A. R. & Karplus, M., Is protein unfolding the reverse of protein folding? A lattice simulation analysis, *J. Mol. Biol.* **292**, 403–419 (1999)
- [103] Micheletti, C., Banavar, J. R. & Maritan, A., Conformations of proteins in equilibrium, *Phys. Rev. Lett.* **87**, 88102 (2001)
- [104] Klimov, D. K. & Thirumalai, D., Multiple protein folding nuclei and transition state ensemble in two-state proteins, *Proteins: Struct. Func. Genet.* **43**, 465–475 (2001)
- [105] Skolnick, J., Kolinski, A., Brooks, C. L., Godzik, A. & Rey, A., A method for predicting protein-structure from sequence, *Curr. Opin. Struct. Biol.* **3**, 414–423 (1993)
- [106] Shakhnovich, E. I., Proteins with selected sequences fold into unique native conformation, *Phys. Rev. Lett.* **72**, 3907–3910 (1994)
- [107] Hao, M. H. & Scheraga, H. A., Monte Carlo simulation of a first-order transition for protein folding, *J. Phys. Chem.* **98**, 4940–4948 (1994)

- [108] Sali, A., Shakhnovich, E. I. & Karplus, M., Kinetics of protein folding. A lattice model study for the requirements for folding to the native state, *J. Mol. Biol.* **235**, 1614–1636 (1994)
- [109] Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I., Impact of local and non-local interactions on thermodynamics and kinetics of protein folding, *J. Mol. Biol.* **252**, 460–471 (1995)
- [110] Kolinski, A., Galazka, W. & Skolnick, J., On the origin of the cooperativity of protein folding: Implications from model simulations, *Proteins* **26**, 271–287 (1996)
- [111] Irbäck, A. & Schwarze, H., Sequence dependence of self-interacting random chains, *J. Phys. A: Math. Gen.* **28**, 2121–2132 (1995)
- [112] Berriz, G. F., Gutin, A. M. & Shakhnovich, E. I., Cooperativity and stability in a Langevin model of proteinlike folding, *J. Chem. Phys.* **106**, 9276–9285 (1997)
- [113] Guo, Z. & III, C. L. Brooks, Thermodynamics of protein folding: a statistical mechanical study of a small all- $\beta$  protein, *Biopolymers* **42**, 745–757 (1997)
- [114] Shea, J. E., Nochomovitz, Y. D., Guo, Z. & III, C. L Brooks, Exploring the space of protein folding Hamiltonians: The balance of forces in a minimalist  $\beta$ -barrel model, *J. Chem. Phys.* **109**, 2895–2903 (1998)
- [115] Klimov, D. K., Newfield, D. & Thirumalai, D., Simulations of beta-hairpin folding confined to spherical pores using distributed computing, *Proc. Natl. Acad. Sci. U. S. A.* **99**, 8019–8024 (2002)
- [116] Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I., Molecular dynamics studies of folding of a protein-like model, *Folding & Design* **3**, 577–587 (1998)
- [117] Nymeyer, H., E.Garcia, A. & Onuchic, J. N., Folding funnels and frustration in off-lattice minimalist protein landscapes, *Proc. Natl. Acad. Sci. U. S. A.* **95**, 5921–5928 (1998)
- [118] Sasai, M., Conformation, energy, and folding ability of selected amino acid sequences, *Proc. Natl. Acad. Sci. U. S. A.* **92**, 8438–8442 (1995)
- [119] Zhou, Y. & Karplus, M., Interpreting the folding kinetics of helical proteins, *Nature* **401**, 400–403 (1999)
- [120] Dokholyan, N. V., Li, L., Ding, F. & Shakhnovich, E. I., Topological determinants of protein folding, *Proc. Natl. Acad. Sci. U. S. A.* **99**, 8637–8641 (2002)
- [121] Chan, H. S. & Dill, K. A., Transition states and folding dynamics of proteins and heteropolymers, *J. Chem. Phys.* **100**, 9238–9257 (1994)
- [122] Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M. & Dobson,

- C. M., Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding, *Nature Struct. Biol.* **6**, 1005–1009 (1999)
- [123] Clarke, J., Cota, E., Fowler, S. B. & Hamill, S. J., Folding studies of immunoglobulin-like beta-sandwich proteins suggest that they share a common folding pathway, *Structure* **7**, 1145–1153 (1999)
- [124] Paci, E., Vendruscolo, M. & Karplus, M., Native and non-native interactions along protein folding and unfolding pathways, *Proteins: Struct. Func. Genet.* **47**, 379–392 (2002)
- [125] Mirny, L. A., Abkevich, V. I. & Shakhnovich, E. I., How evolution makes proteins fold quickly, *Proc. Natl. Acad. Sci. U. S. A.* **95**, 4976–4981 (1998)
- [126] Ptitsyn, O. B. & Ting, K.-L. H., Non-functional conserved residues in globins and their possible role as a folding nucleus, *J. Mol. Biol.* **291**, 671–682 (1999)
- [127] Villegas, V., Martínez, J. C., Avilés, F. X. & Serrano, L., Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain, *J. Mol. Biol.* **283**, 1027–1036 (1998)
- [128] Plaxco, K. W., Simons, K. T. & Baker, D., Contact order, transition state placement and the refolding rates of single domain proteins, *J. Mol. Biol.* **277**, 985–994 (1998)
- [129] Fersht, A. R., Transition-state structure as a unifying basis in protein-folding mechanisms: Contact order, chain topology, stability, and the extended nucleus mechanism, *Proc. Natl. Acad. Sci. U. S. A.* **97**, 1525–1529 (2000)
- [130] Plaxco, K. W., Larson, S., Ruczinski, I., Riddle, D. S., Thayer, E. C., Buchwitz, B., Davidson, A. R. & Baker, D., Evolutionary conservation in protein folding kinetics, *J. Mol. Biol.* **278**, 303–312 (2000)
- [131] Duan, Y. & Kollman, P., Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution, *Science* **282**, 740–743 (1998)
- [132] Smith, S. W., Hall, C. K. & Freeman, B. D., Molecular dynamics for polymeric fluids using discontinuous potentials, *J. Comput. Phys.* **134**, 16–30 (1997)
- [133] Zhou, Y., Karplus, M., Wichert, J. M. & Hall, C. K., Equilibrium thermodynamics of homopolymers and clusters: molecular dynamics and Monte Carlo simulations of system with square-well interactions, *J. Chem. Phys.* **107**, 10691–10708 (1997)
- [134] Dokholyan, N. V., Pitard, E., Buldyrev, S. V. & Stanley, H. E., Glassy behavior of a homopolymer from molecular dynamics simulations, *Phys. Rev. E* **65**, 030801(R) (2002)

- [135] Zhou, Y. & Karplus, M., Folding of a model three-helix bundle protein: A thermodynamic and kinetic analysis, *J. Mol. Biol.* **293**, 917–951 (1999)
- [136] Smith, A. V. & Hall, C. K., Protein refolding versus protein aggregation: computer simulations on an intermediate-resolution protein model, *J. Mol. Biol.* **312**, 187–202 (2001)
- [137] Ding, F., Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I., Molecular dynamics simulation of C-Src SH3 aggregation suggests a generic amyloidogenesis mechanism, *J. Mol. Biol.* , in press (2002)
- [138] Alder, B. J. & Wainwright, T. E., Studies in molecular dynamics. I. General method, *J. Chem. Phys.* **31**, 459–466 (1959)
- [139] Grosberg, A. Yu. & Khokhlov, A. R., *Giant molecules*, Academic Press, Boston, 1997
- [140] Allen, M. P. & Tildesley, D. J., *Computer simulation of liquids*, Clarendon Press, Oxford, 1987
- [141] Rapaport, D. C., *The art of molecular dynamics simulation*, Cambridge University Press, Cambridge, 1997
- [142] Jackson, S. E., How do small single-domain proteins fold?, *Folding & Design* **3**, R81–R91 (1998)
- [143] Doi, M., ed., *Introduction to polymer physics*, Oxford University Press, New York, 1997
- [144] Kabsch, W., A discussion of the solution for the best rotation to relate two sets of vectors, *Acta Cryst.* **A34**, 827–828 (1978)
- [145] Grosberg, A. Yu. & Khokhlov, A. R., *Statistical physics of macromolecules*, AIP Press, New York, 1994
- [146] Vendruscolo, M., Paci, E., Dobson, C. & Karplus, M., Three key residues form a critical contact network in a protein folding transitional state, *Nature* **409**, 641–645 (2001)
- [147] Itzhaki, L. S., Otzen, D. E. & Fersht, A. R., The structure of the transition-state for folding of chymotrypsin inhibitor-2 analyzed by protein engineering methods — evidence for a nucleation-condensation mechanism for protein-folding, *J. Mol. Biol.* **254**, 260–288 (1995)
- [148] Larson, S. M. & Davidson, A. R., The identification of conserved interactions within the SH3 domain by alignment of sequences and structures, *Protein Science* **9**, 2170–2180 (2000)
- [149] Jernigan, R. L. & Bahar, I., Structure-derived potentials and protein simulations, *Curr. Opinion Struc. Biol.* **6**, 195–209 (1996)
- [150] Vendruscolo, M., Dokholyan, N. V., Paci, E. & Karplus, M., A small-world view of the amino

acids that play a key role in protein folding, *Phys. Rev. E* **65**, 061910 (2002)

- [151] Treptow, W. L., Barbosa, M. A. A., Garcia, L. G. & de Araújo, A. F. P., Non-native interactions, contact order and protein folding: A mutational investigation with the hydrophobic model, *Proteins* **49**, 167–180 (2002)

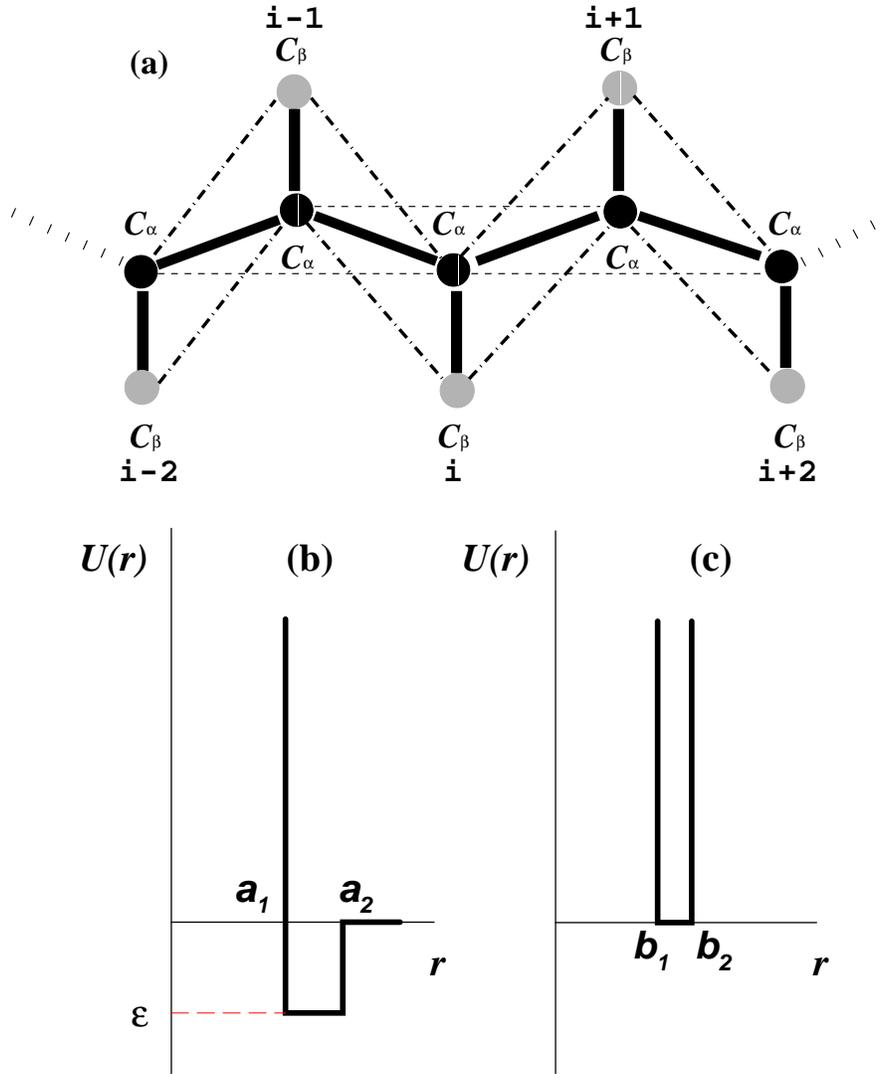


FIG. 1: (a) Schematic diagram of the protein model. Grey spheres represent alpha carbons, black ones represent beta carbons (for Gly, alpha and beta carbons are the same). In the present model only the interaction between side chains are counted, so that the interaction only exists between  $\beta$  carbons, and the  $\alpha$  carbon only plays the role of the backbone. (b,c) The potential of interaction between (b) specific residues; (c) constrained residues.  $a_1$  is the diameter of the hard sphere and  $a_2$  is the diameter of the attractive sphere.  $[b_1, b_2]$  is the interval where residues that are neighbors on the chain can move freely.  $\epsilon$  is negative for native contacts and positive for non-native ones.

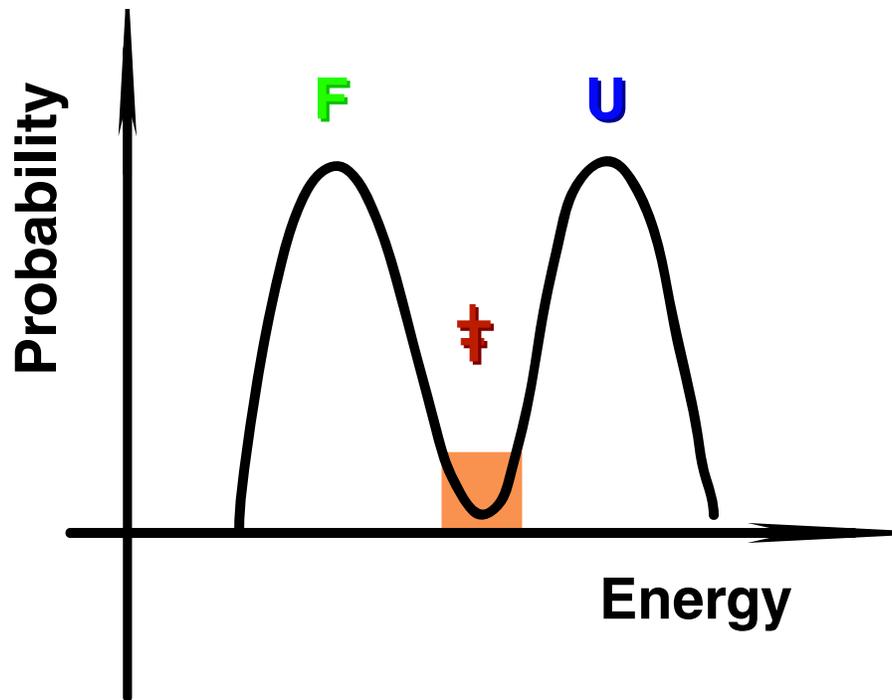


FIG. 2: An illustration of the probability distribution of the potential energies of conformations of two-state proteins at folding transition temperature. The two maxima represent the folded (F) and unfolded (U) conformations, which are separated from each other by low-probability transition states.

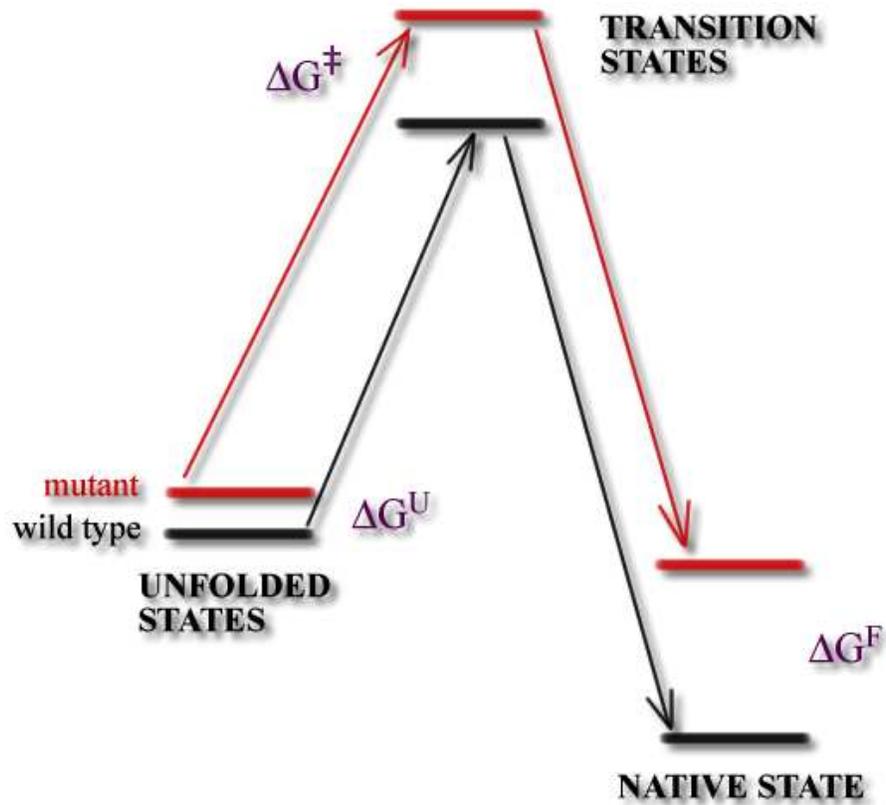


FIG. 3: An illustration of the  $\Phi$ -value analysis for a two-state protein. An amino acid at a specific position is selected in the wild type protein and is mutated to a specific target one. Such mutations affects the free energies of the unfolded, transition, and native states. The extent to which the transition state is affected with respect to the unfolded and native states is measured by  $\Phi$ -values, defined in Eq.(1).

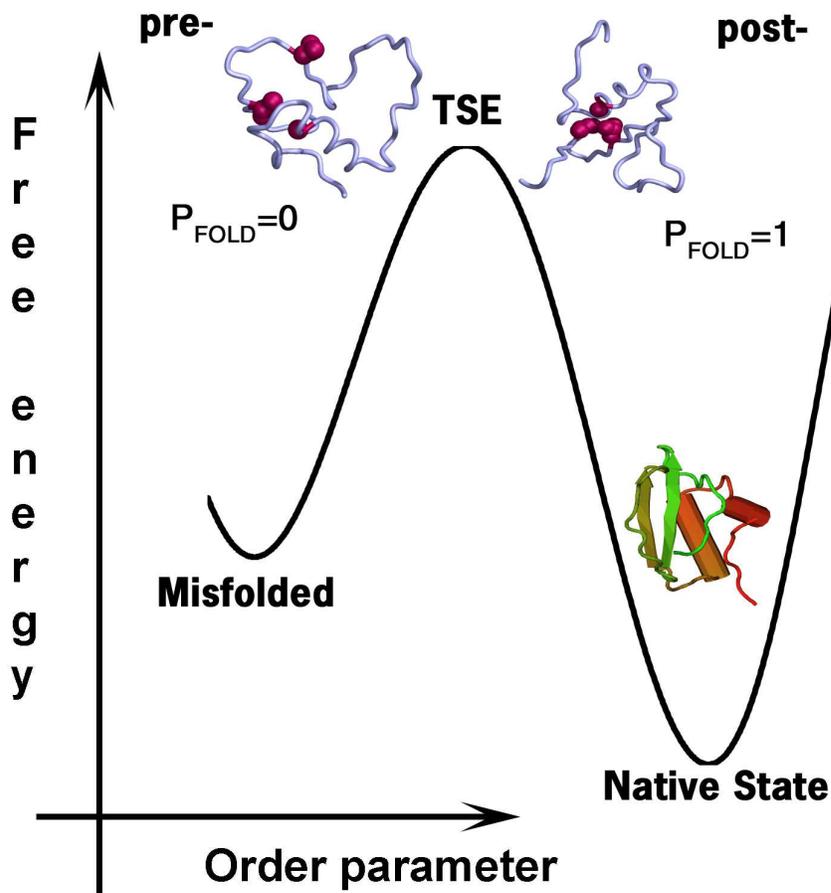


FIG. 4: Two-state protein free energy landscapes are characterized by two distinct minima at the folding transition temperature. One minimum corresponds to misfolded/unfolded set of protein conformations, while the other one corresponds to the native conformation. The two minima are separated by the free energy barrier. The set of conformations at the top of the free energy barrier constitutes the transition state ensemble and is characterized by their probability to rapidly fold to the native state,  $p_{\text{FOLD}} \approx 1/2$ . In pre-transition states, the folding nucleus is not formed, thus the probability to fold of such conformations is close to zero. In post-transition states, the folding nucleus is formed, thus the probability to fold of such conformations is close to 1. The difference in the folding kinetics of pre- and post-transition conformations is drastic, even though their potential energies,  $R_g$ ,  $RMSD$ , and other structural characteristics are close to each other. This difference is exemplified with pre- and post-transition conformations of CI2, obtained by all-atom Monte-Carlo simulations [120]. In pre-transition states the nucleus, A16, L49, and I57 (red beads), is not formed, while in post-transition states the nucleus is intact.

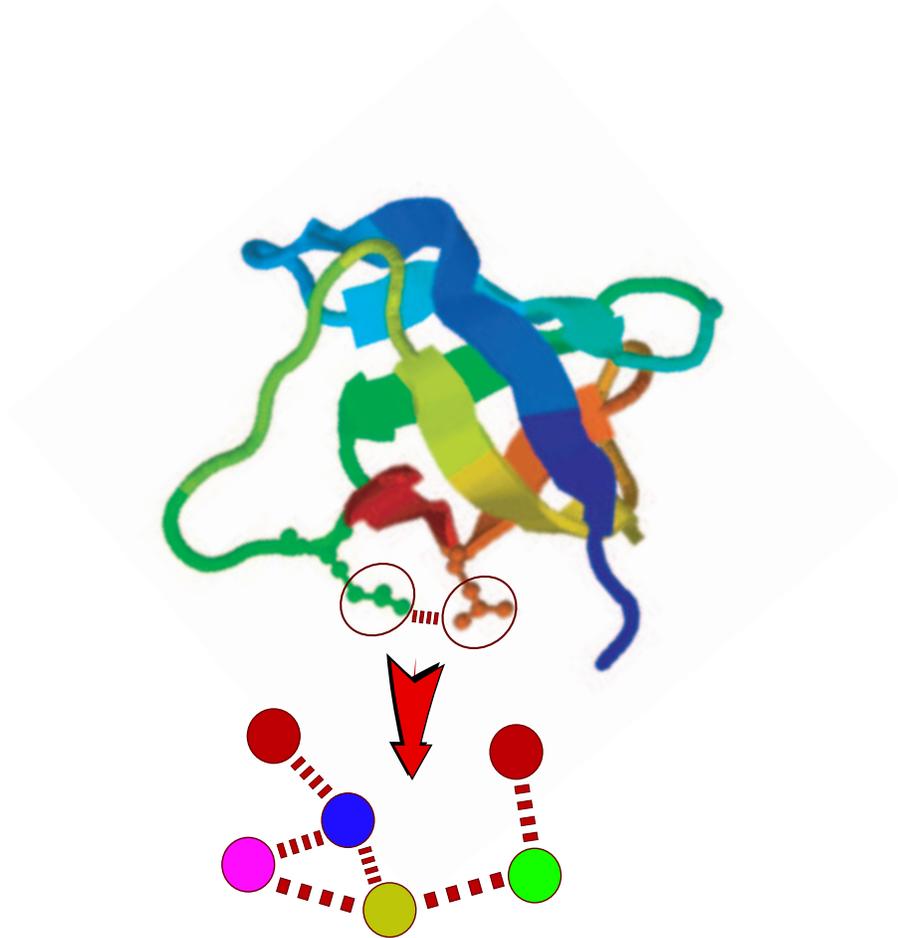


FIG. 5: Constructing protein graphs from protein conformations. Each node corresponds to an amino acid. We draw an edge between any two nodes of a graph if there exists a contact between amino acids, corresponding to these nodes. The contact between two amino acids is defined by the spatial proximity of  $\beta$ -carbons ( $C_\alpha$  for Gly) of these amino acids. The contact distance is taken to be  $8.5\text{\AA}$ .