[25] Identifying Importance of Amino Acids for Protein Folding from Crystal Structures

By Nikolay V. Dokholyan, Jose M. Borreguero, Sergey V. Buldyrev, Feng Ding, H. Eugene Stanley, and Eugene I. Shakhnovich

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 to manipulate protein function is, in turn, crucial for effective drug discovery.

[1] C. B. Anifsen, *Science* **181,** 223 (1973).

[2] H. Taketomi, Y. Ueda, and N. Gō, *Int. J. Pept. Protein Res.* **7,** 445 (1975).

[3] N. Gō, *Annu. Rev. Biophys. Bioeng.* **12,** 183 (1983).

[4] J. D. Bryngelson and P. G. Wolynes, *J. Phys. Chem.* **93,** 6902 (1989).

[5] M. Karplus and E. I. Shakhnovich, *in* "Protein Folding" (T. Creighton, ed.). W. H. Freeman, New York, 1994.

[6] O. B. Ptitsyn, *Protein Eng.* **7,** 593 (1994).

[7] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, *Biochemistry* **33,** 10026 (1994).

[8] E. I. Shakhnovich, V. I. Abkevich, and O. Ptitsyn, *Nature* **379,** 96 (1996).

[9] D. K. Klimov and D. Thirumalai, *Phys. Rev. Lett.* **76,** 4070 (1996).

[10] A. R. Fersht, *Curr. Opin. Struct. Biol.* **7,** 3 (1997).

[11] E. I. Shakhnovich, *Curr. Opin. Struct. Biol.* **7,** 29 (1997).

[12] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, *Annu. Rev. Phys. Chem.* **48,** 545 (1997).

[13] E. I. Shakhnovich, *Fold. Des.* **3,** R45 (1998).

[14] C. Micheletti, J. R. Banavar, A. Maritan, and F. Seno, *Phys. Rev. Lett.* **82,** 3372 (1998).

[15] V. S. Pande, A. Yu Grosberg, D. S. Rokshar, and T. Tanaka, *Curr. Opin. Struct. Biol.* **8,** 68 (1998).

[16] V. P. Grantcharova, D. S. Riddle, J. V. Santiago, and D. Baker, *Nat. Struct. Biol.* **5,** 714 (1998).

[17] J. C. Martinez, M. T. Pissabarro, and L. Serrano, *Nat. Struct. Biol.* **5,** 721 (1998).

[18] H. S. Chan and K. A. Dill, *Proteins Struct. Funct. Genet.* **30,** 2 (1998).

[19] A. F. P. de Araújo, *Proc. Natl. Acad. Sci. USA* **96,** 12482 (1999).

[20] A. R. Dinner and M. Karplus, *J. Chem. Phys.* **37,** 7976 (1999).

[21] B. D. Bursulaya and C. L. Brooks, *J. Am. Chem. Soc.* **121,** 9947 (1999).

[22] B. Nölting and K. Andert, *Proteins Struct. Funct. Genet.* **41,** 288 (2000).

[23] S. B. Ozkan, I. Bahar, and K. A. Dill, *Nat. Struct. Biol.* **8,** 765 (2001).

[24] N. V. Dokholyan, *Recent Res. Dev. Stat. Phys.* **1,** 77 (2001).

[25] A. V. Finkelstein and O. B. Ptitsyn, eds., "Protein Physics: A Course of Lectures." Academic Press, Boston, 2002.

0076-6879/03 $35.00

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 nonhomologous 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 (1) protein function, (2) stability, and (3) folding kinetics—the ability of proteins to rapidly reach their native state.[27,28] Interestingly, function is not conserved among nonhomologous 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 the conservation of amino acids observed in the course of evolution can be explained by pressure to preserve protein stability. Thus, 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 the difficulties and cost of actual experimental studies, it is important to develop rapid computational tools to identify the folding kinetics of a given protein from its crystal structure. The ultimate goal is to be able to predict the protein-folding kinetics of a given protein from its sequence. However, this goal requires the solution of the protein-folding problem (Section II), that is, understanding how a given amino acid sequence folds into a native protein structure. Since protein crystal structures provide invaluable information about amino acid interactions, it is possible to reduce the problem to identifying the folding kinetics of a protein from its structures. The revolution in protein purification and structure-refining methods[29–32] produced a large and constantly growing number of high-quality protein crystal structures. The underlying assumption in these models is that all the information necessary to fold a specific protein is encoded in the protein structure and that the crystal structure can be used as the primary source of information of protein-folding mechanisms. This assumption became the foundation for the development of theoretical and

[26] L. A. Mirny and E. I. Shakhnovich, *J. Mol. Biol.* **291,** 177 (1999).

[27] N. V. Dokholyan and E. I. Shakhnovich, *J. Mol. Biol.* **312,** 289 (2001).

[28] N. V. Dokholyan, L. A. Mirny, and E. I. Shakhnovich, *Physica A* **314,** 600 (2002).

[29] J. Drenth, ''Principles of Protein X-Ray Crystallography.'' Springer-Verlag, New York, 1994

[30] Macromolecular crystallography, Part A. in C. W. Carter, Jr. and R. M. Sweet, eds., *Methods Enzymol.* **276,** (1997).

[31] C. W. Carter, Jr. and R. M. Sweet, eds., *Methods Enzymol.* **277,** (1997).

[32] J. Roach and C. W. Carter, Jr., *Acta Crystallogr. A* **59,** 273–280 (2003).

computational models of protein-folding mechanisms.[4,33–46] Surprisingly, such an approach has already yielded promising and robust results, which are summarized in this chapter.

Here we present an overview of computational techniques for reconstructing the folding mechanisms of proteins from their crystal structures. We also describe methods that we have validated on the Src SH3 domain, a 56-amino acid protein, studied in detail in experiments[16,22,43,47–59] and molecular dynamics simulations.[45,60–63] We describe a new protein model

[33] H. S. Chan and K. A. Dill, *Annu. Rev. Biochem.* **20,** 447 (1991).

[34] A. M. Gutin and E. I. Shakhnovich, *J. Chem. Phys.* **98,** 8174 (1993).

[35] C. J. Camacho and D. Thirumalai, *Proc. Natl. Acad. Sci. USA* **90,** 6369 (1993).

[36] J. D. Bryngelson, *J. Chem. Phys.* **100,** 6038 (1994).

[37] V. S. Pande, A. Y. Grosberg, and T. Tanaka, *Proc. Natl. Acad. Sci. USA* **91,** 12972 (1994).

[38] A. J. Li and V. Daggett, *Proc. Natl. Acad. Sci. USA* **91,** 10430 (1994).

[39] H. Li, C. Tang and N. S. Wingreen, *Proc. Natl. Acad. Sci. USA* **95,** 4987 (1998).

[40] V. Munoz and W. A. Eaton, *Proc. Natl. Acad. Sci. USA* **96,** 11311 (1999)

[41] O. V. Galzitskaya and A. V. Finkelstein, *Proc. Natl. Acad. Sci. USA* **96,** 11299 (1999).

[42] E. Alm and D. Baker, *Proc. Natl. Acad. Sci. USA* **96,** 11305 (1999).

[43] R. Guerois and L. Serrano, *J. Mol. Biol.* **304,** 967 (2000).

[44] H. Nymeyer, N. D. Socci, and J. N. Onuchic, *Proc. Natl. Acad. Sci. USA* **97,** 634 (2000).

[45] C. Clementi, H. Nymeyer, and J. N. Onuchic, *J. Mol. Biol.* **278,** 937 (2000).

[46] N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, *J. Mol. Biol.* **296,** 1183 (2000).

[47] A. P. Combs, T. M. Kapoor, S. Feng, and J. K. Chen, *J. Am. Chem. Soc.* **118,** 287 (1996).

[48] S. Feng, J. K. Chen, H. Yu, J. A. Simons, and S. L. Schreiber, *Science* **266,** 1241 (1994).

[49] S. Feng, C. Kasahara, R. J. Rickles, and S. L. Schreiber, *Proc. Natl. Acad. Sci. USA* **92,** 12408 (1995).

[50] H. Yu, M. K. Rosen, T. B. Shin, C. Seidel-Dugan, and J. S. Brugde, *Science* **258,** 1665 (1992).

[51] V. P. Grantcharova and D. Baker, *Biochemistry* **36,** 15685 (1997).

[52] D. S. Riddle, V. P. Grantcharova, J. V. Santiago, E. Alm, I. Ruczinski, and D. Baker, *Nat. Struct. Biol.* **6,** 1016 (1999).

[53] A. R. Viguera, J. C. Martinez, V. V. Filimonov, P. L. Mateo, and L. Serrano, *Biochemistry* **33,** 10925 (1994).

[54] A. R. Viguera, L. Serrano, and M. Wilmanns, *Nat. Struct. Biol.* **10,** 874 (1996).

[55] J. C. Martinez, A. R. Viguera, R. Berisio, M. Wilmanns, P. L. Mateo, V. V. Filimonov, and L. Serrano, *Biochemistry* **38,** 549 (1999).

[56] S. Knapp, P. T. Mattson, P. Christova, K. D. Berndt, A. Karshikoff, M. Vihinen, C. I. Smith, and R. Ladenstein, *Proteins Struct. Funct. Genet.* **23,** 309 (1998).

[57] Y. -K. Mok, E. L. Elisseeva, A. R. Davidson, and J. D. Forman-Kay, *J. Mol. Biol.* **307,** 913 (2001).

[58] J. G. B. Northey, A. A. Di Nardo and A. R. Davidson, *Nat. Struct. Biol.* **9,** 126 (2002).

[59] J. G. B. Northey, K. L. Maxwell, and A. R. Davidson, *J. Mol. Biol.* **320,** 389 (2002).

[60] J. Gsponer and A. Catfilsch, *J. Mol. Biol.* **309,** 285 (2001).

[61] C. Clementi, P. A. Jennings, and J. N. Onuchic, *J. Mol. Biol.* **311,** 879 (2001).

[62] J. M. Borreguero, N. V. Dokholyan, S. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, *J. Mol. Biol.* **318,** 863 (2002).

[63] F. Ding, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, *Biophys. J.* **83,** 3525 (2002).

and show that the thermodynamics of Src SH3 from molecular dynamics simulations is consistent with that observed experimentally. 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.[7,13,15,46,64]

How Do Proteins Fold?

Levinthal "Paradox"

In 1968, Levinthal formulated a simple argument that points out the nonrandom character of protein-folding kinetics,[65] 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, and 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 on the order of magnitude of 1 *ps*, 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 range of 1 *ms*–1 *s*. Thus, there is a specific mechanism due to which proteins avoid most conformations en route to their native state.

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."[12,66] Various measures have been proposed to determine whether a protein sequence qualifies as a two-state folder, either relying on kinetic[67] or thermodynamic[9,68] properties.

The free energy landscape of the two-state proteins at physiological temperatures is characterized by two deep minima.[7,11,15,41,53,69,70] One

[64] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, *Fold. Des.* **3,** 183 (1998).

[65] C. Levinthal, *J. Chim. Phys.* **65,** 44 (1968).

[66] J. N. Onuchi, H. Nymeyer, A. E. Garcia, J. Chahine, and N. D. Socci, *Adv. Protein Chem.* **53,** 87 (2000).

[67] J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, *Proteins Struct. Funct. Genet.* **21,** 167 (1995).

[68] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, *Folding Design* **1,** 221 (1996).

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 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 and unfolded states coexist with equal probabilities. The two minima are separated by a free energy barrier. The set of conformations that belong to the top of this barrier, having maximal values of free energy, is called the *transition state ensemble* (TSE).

At equilibrium, the probability of observing a conformation with free energy, ΔG, 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 the free energies of native and unfolded or misfolded ensembles are equal, the probability of existing 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 probability = 1/2. So, the TSE is characterized by the probability of the conformations to directly reach the native state without unfolding equal to 1/2.[62,63,71]

The questions then concern which conformations belong to the top of the free energy barrier, and whether there any specific mechanisms that are responsible for the rapid folding transition. Numerous folding scenarios have been proposed to answer these questions.[6,13,42,72–77] The mechanism that we advocate in this chapter 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

[69] S. E. Jackson, N. Elmasry, and A. R. Fersht, *Biochemistry* **32,** 11270 (1993).

[70] J. P. K. Doye and D. J. Wales, *J. Chem. Phys.* **105,** 8428 (1996).

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

[72] O. B. Ptitsyn, *Dokl. Acad. Nauk.* **210,** 1213 (1973).

[73] P. S. Kim and R. L. Baldwin, *Annu. Rev. Biochem.* **59,** 631 (1994).

[74] M. Karplus and D. L. Weaver, *Protein Sci.* **3,** 650 (1994).

[75] D. B. Wetlaufer, *Proc. Natl. Acad. Sci. USA* **70,** 691 (1973).

[76] D. B. Wetlaufer, *Trends Biochem. Sci.* **15,** 414 (1990).

[77] O. B. Ptitsyn, *Nat. Struct. Biol.* **3,** 488 (1996).

we must determine the nucleus and the TSE of a protein. Next, we describe a protein model that we use in molecular dynamics simulations (Fig. 1).

Protein Engineering Experiments

A way to test the importance of amino acids in experiments was proposed by Fersht and co-workers.[78,79] The method, called *protein engineering* or Φ value analysis, is based on the engineering of a mutant protein with the amino acids under consideration replaced by other ones. The value of the free energy difference between the wild-type and the mutant proteins is measured at the transition state $\Delta G^{\ddagger}$, folded state ΔG^{F}, and unfolded states ΔG^{U} (Fig. 3). Φ values are defined as

$$\Phi = \frac{\Delta G^{\ddagger} - \Delta G^{\mathrm{U}}}{\Delta G^{\mathrm{F}} - \Delta G^{\mathrm{U}}} \tag{1}$$

Φ 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. Φ values are close to unity for those amino acids whose substitution affects the transition states to the same extent as the folded states. Thus, these amino acids are most important for protein-folding kinetics.

Developments in Determination of Protein-Folding Kinetics

Developments in protein purification and structure-refining methods[29] have led to publication of high-resolution protein crystal structures. This set of data boosted theoretical studies of protein folding beyond the general heteropolymer models.[4,33–37] Early studies targeting important amino acids for protein dynamics applied the available crystal structure data in two different approaches: structures were used (1) as reference states (decoys) for theoretical predictions,[38,80–86] and (2) as a source of dynamic

[78] A. Matouschek, J. T. Kelis, Jr., L. Serrano, and A. R. Fersht, *Nature* **342,** 122 (1989).

[79] A. Matouschek, J. T. Kelis, Jr., L. Serrano, M. Bycroft, and A. R. Fersht, *Nature* **346,** 440 (1990).

[80] E. M. Boczko and C. L. Brooks, *Science* **269,** 393 (1995).

[81] V. Daggett, A. J. Li, L. S. Itzhaki, D. E. Otzen, and A. R. Fersht, *J. Mol. Biol.* **257,** 430 (1996).

[82] T. Lazaridis and M. Karplus, *Science* **278,** 1928 (1997).

[83] F. B. Sheinerman and C. L. Brooks III, *Proc. Natl. Acad. Sci. USA* **95,** 1562 (1998).

[84] S. L. Kazmirski and V. Daggett, *J. Mol. Biol.* **284,** 793 (1998).

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

[86] N. A. Marti-Renom, R. H. Stote, E. Querol, F. X. Aviles, and M. Karplus, *J. Mol. Biol.* **284,** 145 (1998).

information.[87–91] Studies relied on the developed theoretical framework[92] that explained folding of relatively small proteins as a chemical reaction between two sets of species—folded and denaturated protein states, separated by transition states and possibly by 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 limitations and the inaccuracies in the interatomic force field[93] forced all-atom-folding simulations to be performed under extreme conditions favoring denaturation, typically high temperatures.[38,80–85] This approach assumes that protein folding can be described by running the unfolding simulation backward 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 kilocalories per mole,[94] 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.[95]

In simulations, Daggett *et al.*[38,81,84] 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 proteins, 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 Φ values.

The issue of the limited statistical significance of the results,[38,81,84] due to a small number of unfolding simulations, was addressed by Lazaridis and

[87] C. Wilson and S. Doniach, *Proteins* **6,** 193 (1989).

[88] J. Skolnick and A. Kolinski, *Science* **250,** 1121 (1990).

[89] K. A. Dill, K. M. Fiebig, and H. S. Chan, *Proc. Natl. Acad. Sci. USA* **90,** 1942 (1993).

[90] A. Kolinski and J. Skolnick, *Proteins Struct. Funct. Genet.* **18,** 338 (1994).

[91] M. Vieth, A. Kolinski, C. L. Brooks, and J. Skolnick, *J. Mol. Biol.* **251,** 448 (1995).

[92] E. I. Shakhnovich and A. M. Gutin, *Biophys. Chem.* **34,** 187 (1989).

[93] W. Wang, O. Donini, C. M. Reyes, and P. A. Kollman, *Annu. Rev. Biophys. Biophys. Struct.* **30,** 211 (2001).

[94] T. Creighton, "Proteins: Structures and Molecular Properties," 2nd Ed. W. H. Freeman, New York, 1993.

[95] A. V. Finkelstein, *Protein Eng.* **10,** 843 (1997).

Karplus,[82] who performed a larger series of unfolding simulations starting from conformations slightly different from the initial crystal structure. A wealth of simulations allowed those authors to extract the common set of key interactions and to 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.[80,83] 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.

All-atom simulations were also used to increase the efficiency of protein-engineering experiments in a self-consistent experimental and computational approach toward determination of the TSE.[85] This method is most useful for proteins for which only a small fraction of the residues play a key role. Such a method may also serve as a refining tool of the protein-engineering results.

Protein databases[96] of crystal structures have been widely used as a source of dynamic information with application to folding simulations. In their pioneer study, Wilson and Doniach[87] computed effective pairwise amino acid contact potentials from the frequencies of spatial proximities between pairs of amino acids obtained in the database of structures. The authors used these potentials to reproduce, with modest success, the folding process of a one-atom per residue model of crambin on a square lattice.[97] Skolnick and Kolinski[88] developed a statistical potential using two-atom representation of apoplastocyanin.[98] Folding simulations on a finer lattice than that used in previous studies allowed the authors to fold a model protein with a root–mean–square deviation (RMSD) of 6 Å with respect to the crystal structure. However, the propensity of the amino acids to adopt a specific crystal structure prevented the authors from generalizing the applicability of the model to more than one protein at a time.

Kolinsky and Skolnick[90] extended the original model[88] 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 the authors to describe folding processes with moderate success. Vieth *et al.*[91] used a similar model to study the aggregation kinetics of the GCN4 leucine zipper[99] into dimers, trimers, and

[96] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, *Nucleic Acids Res*. **28,** 235 (2000).

[97] M. M. Teeter, S. M. Roe, and N. H. Heo, *J. Mol. Biol.* **230,** 292 (1993).

[98] T. P. Garrett, D. J. Clingeleffer, J. M. Guss, S. J. Rogers, and H. C. Freeman, *J. Biol. Chem*. **259,** 2822 (1984).

[99] E. K. O'Shea, J. D. Klemm, P. S. Kim, and T. Alber, *Science* **254,** 539 (1991).

tetramers. The authors identified the amino acids that regulated the aggregation kinetics, in accordance with experiments. A systematic study[100] 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 with *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.*[89]in their study of the folding mechanisms of crambin and chymotrypsin inhibitor.[101] Dill *et al*. assigned attractive interactions between all pairs of hydrophobic amino acids, neglecting other amino acid interactions. The folding dynamics was implemented through a sequence of folding events in which hydrophobic contacts act as constraints that bring other contacts into spatial proximity. The authors found that 1 in 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 identifying the important amino acids for protein folding have attracted interest.[38,39] These methods use the crystal structure as the reference state. Starting from the crystal structure, temperature-induced unfolding[38,39] in all-atom molecular dynamics simulations with explicit solvent molecules have been applied to study the transition states. However, the limitation of the computational ability of traditional molecular dynamics algorithms only enables one to sample over several unfolding trajectories from the folded state. Thus, this technique can only capture one or a few transition state conformations instead of a statistically significant ensemble. Moreover, derivation of a folding transition state ensemble from high-temperature unfolding may be problematic in some cases due to possible significant differences between the high-temperature free energy landscape and the free energy landscape of a protein at physiological temperatures.[95,102]

[100] D. Thomas, G. Casari, and C. Sander, *Protein Eng*. **9,** 941 (1996).

[101] C. A. McPhalen and M. N. James, *Biochemistry* **26,** 261 (1987).

[102] A. R. Dinner and M. Karplus, *J. Mol. Biol*. **292,** 403 (1999).

Alternative theoretical approaches[40–43] have been proposed to predict the transition states in protein folding and have obtained significant correlations with experimental Φ values for several proteins. However, each of these models involves drastic assumptions. For example, each amino acid 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 an effective dynamic algorithm, simplified protein models with a crystal structure-based interaction potential[44,45,103] have been applied to study folding kinetics. The principal difficulty in the kinetics studies is the classification of various protein conformations, that is, the knowledge of the *reaction coordinate*—a parameter that can uniquely identify the position of a protein conformation on a folding landscape with respect to the native state. The fraction of native contacts Q[44,45] 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,[7,71,104] and the principal difficulty in identifying the folding reaction coordinate from crystal structures is in uncovering the relationship between protein-folding thermodynamics and kinetics, that is, how much kinetic information we can obtain about protein folding barriers from equilibrium sampling of folding trajectories.

Protein-Folding Kinetics from Discrete Molecular Dynamics Simulations

Protein Model

The problem of protein modeling in simulations is as complex as the protein-folding problem itself. Such complexity often makes brute force approaches of all-atom simulations impractical. Lattice models[7,35,70,105–110] became popular due to their ability to reproduce a significant amount of

[103] C. Micheletti, J. R. Banavar, and A. Maritan, *Phys. Rev. Lett.* **87,** 88102 (2001).

[104] D. K. Klimov and D. Thirumalai, *Proteins Struct. Funct. Genet.* **43,** 465 (2001).

[105] J. Skolnick, A. Kolinski, C. L. Brooks, A. Godzik, and A. Rey, *Curr. Opin. Struct. Biol.* **3,** 414 (1993).

[106] E. I. Shakhnovich, *Phys. Rev. Lett.* **72,** 3907 (1994).

[107] M. H. Hao and H. A. Scheraga, *J. Phys. Chem.* **98,** 4940 (1994).

[108] A. Sali, E. I. Shakhnovich, and M. Karplus, *J. Mol. Biol.* **235,** 1614 (1994).

[109] V. I. Abkevich, A. M. Gutin, and E. I. Shakhnovich, *J. Mol. Biol.* **252,** 460 (1995).

[110] A. Kolinski, W. Galazka, and J. Skolnick, *Proteins* **26,** 271 (1996).

folding events in a reasonable computational time. However, the role of topology (common structural features) in determining the folding nucleus requires study beyond lattice models, which impose unphysical constraints on the protein degrees of freedom. Simplified off-lattice models[45,46,111–115] are a compromise between lattice and all-atom models. They can be regarded as the first step into modeling the realistic 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 beads-on-a-string model is that its chain flexibility is higher than that observed in real proteins, so, as a result, the protein model folding kinetics is often altered, due to 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 Ding *et al.*[63] we model a protein by beads representing C_α and C_β (Fig. 1A). There are four types of bonds: (1) covalent bonds between $C_{\alpha i}$ and $C_{\beta i}$, (2) peptide bonds between $C_{\alpha i}$ and $C_{\alpha(i\ \ 1)}$, (3) effective bonds between $C_{\beta i}$ and $C_{\alpha(i\ \ 1)}$, and (4) effective bonds between $C_{\alpha i}$ and $C_{\alpha(i\ \ 2)}$. To determine the effective bond length, we calculate the average and the standard deviation of distances between carbon pairs of types 3 and 4 for 10^3 representative globular proteins obtained from the Protein Data Bank.[96] We find that the average distances are 4.7 and 6.2 Å for type 3 and type 4 bonds, respectively. The ratio σ of the standard deviation to the average for bond types 3 and 4 are 0.036 and 0.101, respectively. The standard deviation of bond type 4 is larger than that of bond type 3 because it is related to the angle of two consecutive peptide bonds. Thus, the bond lengths of type 4 fluctuate more than those of type 3. 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):

[111] A. Irbäck and H. Schwarze, *J. Phys. A Math. Gen*. **28,** 2121 (1995).

[112] G. F. Berriz, A. M. Gutin, and E. I. Shakhnovich, *J. Chem. Phys*. **106,** 9276 (1997).

[113] Z. Guo and C. L. Brooks III, *Biopolymers* **42,** 745 (1997).

[114] J. E. Shea, Y. D. Nochomovitz, Z. Guo, and C. L. Brooks III, *J. Chem. Phys*. **109,** 2895 (1998).

[115] D. K. Klimov, D. Newfield, and D. Thirumalai, *Proc. Natl. Acad. Sci. USA* **99,** 8019 (2002).

[116] N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, *Folding Design* **3,** 577 (1998).

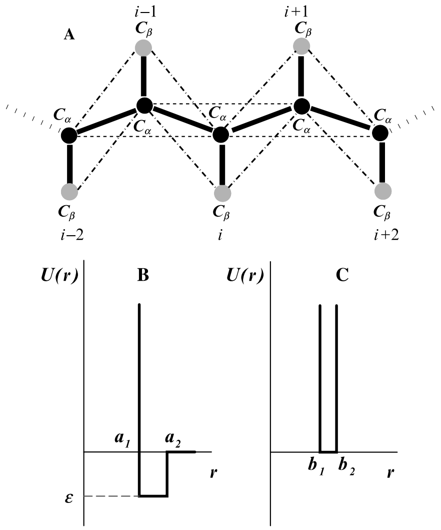

FIG. 1. (A) Schematic diagram of the protein model. Grey spheres represent α carbons, black ones represent β carbons (for Gly, α and β carbons are the same). In the present model only the interactions between side chains are counted, so that the interaction exists only between β carbons, and the α carbon plays only the role of the backbone. (B and 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. ε is negative for native contacts and positive for nonnative ones.

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

where D_{ij} is the distance between atoms i and j in the native state, $\sigma = 0.0075$ for a bond of type 1, $\sigma = 0.02$ for a bond of type 2, $\sigma = 0.036$ for a bond of type 3, and $\sigma = 0.101$ for a bond of type 4. 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[45,114,117] 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.

[117] H. Nymeyer, A. E.Garcia, and J. N. Onuchic, *Proc. Natl. Acad. Sci. USA* **95,** 5921 (1998).
[118] M. Sasai, *Proc. Natl. Acad. Sci. USA* **92,** 8438 (1995).

We use a modified Gō model similar to one described in Dokholyan *et al.*,[116] in which interactions are determined by the native structure of proteins. In our model, only C_β atoms that are not nearest neighbors along the chain interact with each other. The cutoff distance between C_β atoms is chosen to be 7.5 Å. The Gō model has been widely applied to study various aspects of protein-folding thermodynamics and kinetics.[24,42,46,62,63,119,120]

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: (1) the unique and stable native state; (2) a cooperative folding transition resembling a first-order phase transition. Importantly, protein sequences with amino acids represented by only two or three types showed a relatively slow decrease in potential energy at T_F 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 nonspecific interactions to the Gō model resulted in analogous trapping[114]; and (3) 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 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 Gō 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, conclusive proof of such an assumption can be obtained either in simulations that do not use the Gō model, or in experiments that compare folding pathways of analogs—proteins with nonhomologous sequences that fold into similar conformations. The dominating role of topology in defining folding mechanisms 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

[119] Y. Zhou and M. Karplus, *Nature* **401,** 400 (1999).

[120] N. V. Dokholyan, L. Li, F. Ding, and E. I. Shakhnovich, *Proc. Natl. Acad. Sci. USA* **99,** 8637 (2002).

[121] H. S. Chan and K. A. Dill, *J. Chem. Phys.* **100,** 9238 (1994).

[122] F. Chiti, N. Taddei, P. M. White, M. Bucciantini, F. Magherini, M. Stefani, and C. M. Dobson, *Nat. Struct. Biol.* **6,** 1005 (1999).

[123] J. Clarke, E. Cota, S. B. Fowler, and S. J. Hamill, *Structure* **7,** 1145 (1999).

[124] E. Paci, M. Vendruscolo, and M. Karplus, *Proteins Struct. Funct. Genet.* **47,** 379 (2002).

folding nucleus. This finding was further corroborated in Mirny *et al.*,[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 amino acid 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 a common folding nucleus for proteins representing that fold. Experimentally, a common folding nucleus was found in α/β plait proteins that have no sequence homology.[122,127] Other works provided support for the important role of protein topology in its folding kinetics.[45,122,128–130]

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 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] A detailed description of the algorithm can be found in Refs. [138–141]

[125] L. A. Mirny, V. I. Abkevich, and E. I. Shakhnovich, *Proc. Natl. Acad. Sci. USA* **95,** 4976 (1998).

[126] O. B. Ptitsyn and K. -L. H. Ting, *J. Mol. Biol.* **291,** 671 (1999).

[127] V. Villegas, J. C. Martínez, F. X. Avilés, and L. Serrano, *J. Mol. Biol.* **283,** 1027 (1998).

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

[129] A. R. Fersht, *Proc. Natl. Acad. Sci. USA* **97,** 1525 (2000).

[130] K. W. Plaxco, S. Larson, I. Ruczinski, D. S. Riddle, E. C. Thayer, B. Buchwitz, A. R. Davidson, and D. Baker, *J. Mol. Biol.* **278,** 303 (2000).

[131] Y. Duan and P. Kollman, *Science* **282,** 740 (1998).

[132] S. W. Smith, C. K. Hall, and B. D. Freeman, *J. Comput. Phys.* **134,** 16 (1997).

[133] Y. Zhou, M. Karplus, J. M. Wichert, and C. K. Hall, *J. Chem. Phys.* **107,** 10691 (1997).

[134] N. V. Dokholyan, E. Pitard, S. V. Buldyrev, and H. E. Stanley, *Phys. Rev. E* **65,** 030801(R) (2002).

[135] Y. Zhou and M. Karplus, *J. Mol. Biol.* **293,** 917 (1999).

[136] A. V. Smith and C. K. Hall, *J. Mol. Biol.* **312,** 187 (2001).

[137] F. Ding, N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich, *J. Mol. Biol.* **324,** 851 (2002).

[138] B. J. Alder and T. E. Wainwright, *J. Chem. Phys.* **31,** 459 (1959).

[139] A. Y. Grosberg and A. R. Khokhlov, "Giant Molecules." Academic Press, Boston, 1997.

[140] M. P. Allen and D. J. Tildesley, "Computer Simulation of Liquids." Clarendon Press, Oxford, 1987.

[141] D. C. Rapaport, "The Art of Molecular Dynamics Simulation." Cambridge University Press, Cambridge, 1997.

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. The temperature is measured in units of ε/k_B, where ε is the typical interaction strength between pairs of amino acids and k_B is the Boltzmann constant. The time step is equal to the shortest time between two consecutive collisions between any two particles in the system.

Folding Thermodynamics

To test whether our models faithfully reproduce the experimentally observed[16,52,53,142] thermodynamic and kinetic properties of the SH3 domain, Ding *et al.*[63] and Borreguero *et al.*[62] 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 Å. 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 Dokholyan *et al.*[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 the C-Src SH3 domain.[16,52,53,142]

Protein-Folding Kinetics

Identifying Folding Nucleus. A method to identify the protein-folding nucleus from equilibrium trajectories was proposed in Dokholyan *et al.*[116] and later used on SH3 domain proteins.[62,63] The idea is to study ensembles of conformations that have a specific history and future. For example,

[142] S. E. Jackson, *Folding Design* **3,** R81 (1998).

[143] M. Doi, ed., "Introduction to Polymer Physics." Oxford University Press, New York, 1997.

[144] W. Kabsch, *Acta Crystallogr. A* **34,** 827 (1978).

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, to determine the nucleus, we 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[46] and later, in Ding *et al.*[63] to determine the folding nucleus of Src SH3 domain protein. For the 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 the 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 the 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 a control, we test whether 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 the case of a homopolymer, it reduces the entropy of conformational space the most.[145] We find that fixation of T9–S64 does not

[145] A. Y. Grosberg and A. R. Khokhlov, ‘‘Statistical Physics of Macromolecules.’’ AIP Press, New York, 1994.

significantly affect the distribution of energy states, indicating that formation of an arbitrary contact is not a sufficient condition to bias the protein conformation toward its native state.

Identifying 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 whether selected conformations belong to the transition state ensemble by computing its probability to fold, p_{FOLD}.[71] To determine p_{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_{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_{FOLD} values close to 1/2.

We study three types of conformations: (1) UU, (2) FF, and (3) UF. The latter is a set of conformations that originates 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 but lower than the average energy of unfolded states, so that their potential energy has 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_{FOLD} to be close to zero. In FF conformations, the nucleus is not present, and since there is little chance that it will be disrupted, we expect p_{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_{FOLD} to be close to 1/2.

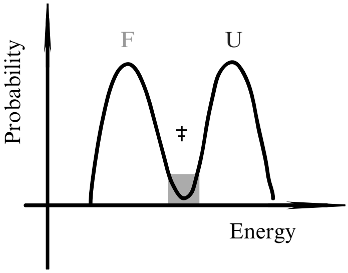

FIG. 2. An illustration of the probability distribution of the potential energies of conformations of two-state proteins at folding transition temperature. The two maxima represent the folded (F) and unfolded (U) conformations, which are separated from each other by low-probability transition states.

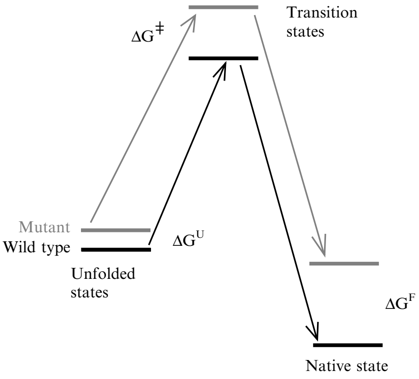

FIG. 3. An illustration of the *Φ* value analysis for a two-state protein. An amino acid at a specific position is selected in the wild-type protein and is mutated to a specific target. Such mutations affect the free energies of the unfolded, transition, and native states. The extent to which the transition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

In Refs. 62 and 63 we show that, in fact, p_{FOLD} is close to zero for the ensemble of UU conformations. p_{FOLD} is approximately unity for the ensemble of FF conformations. Only for the ensemble of UF conformations do we find that p_{FOLD} is close to 1/2. Thus, the set of UF conformations represents 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 that we observe difference in p_{FOLD} values, even though their energetic (potential energy) and structural (RMSD, R_g) characteristics are close to each other.

Virtual Screening Method. We use a technique similar to experimental *Φ* 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:

$$\Delta G_x = -kT \ln \langle \exp(-\Delta E/kT) \rangle_x \tag{3}$$

Here x denotes a state ensemble (folded, F; unfolded, U; and transition, ‡), ΔE is the change in potential energy due to the mutation (details in Ref.[45]), and the average $\langle \ldots \rangle_x$ is taken over all conformations of unfolded, transition, and folded state ensembles. We compute[45]

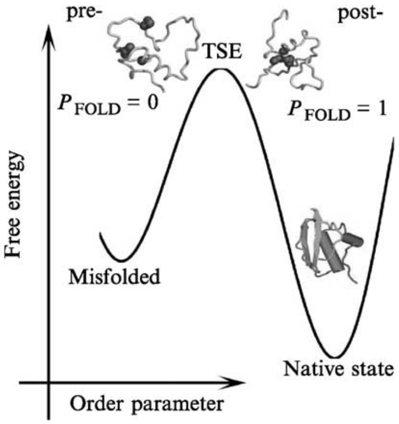

FIG. 4. Two-state protein free energy landscapes are characterized by two distinct minima at the folding transition temperature. One minimum corresponds to a misfolded/unfolded set of protein conformations, while the other corresponds to the native conformation. The two minima are separated by the free 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 pretransition states, the folding nucleus is not formed, and thus the probability to fold of such conformations is close to zero. In posttransition states, the folding nucleus is formed, and thus the probability to fold of such conformations is close to 1. The difference in the folding kinetics of pre- and posttransition 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 posttransition conformations of CI2, obtained by all-atom Monte Carlo simulations.[120] In pretransition states the nucleus, A16, L49, and 157 (beads), is not formed, while in posttransition states the nucleus is intact.

$$\Phi = \frac{\ln\langle \exp(-\Delta E/kT)\rangle_{\ddagger} - \ln\langle \exp(-\Delta E/kT)\rangle_{\mathrm{U}}}{\ln\langle \exp(-\Delta E/kT)\rangle_{F} - \ln\langle \exp(-\Delta E/kT)\rangle_{\mathrm{U}}} \tag{4}$$

The Φ values in our analysis are determined using the free energy relationship of Eq. (3), which 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 Φ value used in more recent work[146] as proportional to the number of contacts a residue makes in the TSE, the correlation coefficient between theoretical and experimental Φ values is reduced to 0.27 from approximately 0.6 (Section V.D.4). An approximation to the Φ value, the difference between the average number

[146] M. Vendruscolo, E. Paci, C. Dobson, and M. Karplus, *Nature* **409,** 641 (2001).

of contacts residues form in the TSE and in unfolded states, $\Phi \approx (\langle N_i \rangle_{\ddagger} - \langle N_i \rangle_{\mathrm{U}}) / (\langle N_i \rangle_{\mathrm{F}} - \langle N_i \rangle_{\mathrm{U}})$, provides a better correlation coefficient between predicted and experimentally observed Φ values (0.48) than does the approximation of Vendruscolo *et al.*[146] The reason that a thermodynamic definition of the Φ value yields better agreement with experiments can be inferred from a ΔG plot,[63] which shows that $\Delta G^{\mathrm{F}} - \Delta G^{\mathrm{U}}$ 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.

Comparing Simulations with Experiments. In Ding *et al.*,[63] Φ values are computed using the *virtual screening method* and the comparison with experimental Φ values[16, 52] 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 156. In general, the majority of the residues from that list have high experimental Φ 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 Φ value—1.2. Notable exceptions are L24, W43, and G54, which have Φ 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.[62,148]

Role of Protein Topology

En route to the native state, at the transition states a protein loses its entropy by forming a specific nucleus (see Fig. 4). Entropically and energetically pre- and posttransition states—conformations with p_{FOLD} approximately zero and unity correspondingly—are indistinguishable. In fact, in Dokholyan *et al.*,[120] pre- and posttransition sets of conformations were selected for SH3 and C12 proteins. Both pre- and posttransition states had similar structural and energetic properties. The question then is: "What distinguishes pre- and posttransition states?"

[147] L. S. Itzhaki, D. E. Otzen and A. R. Fersht, *J. Mol. Biol.* **254,** 260 (1995).

[148] S. M. Larson and A. R. Davidson, *Protein Sci.* **9,** 2170 (2000).

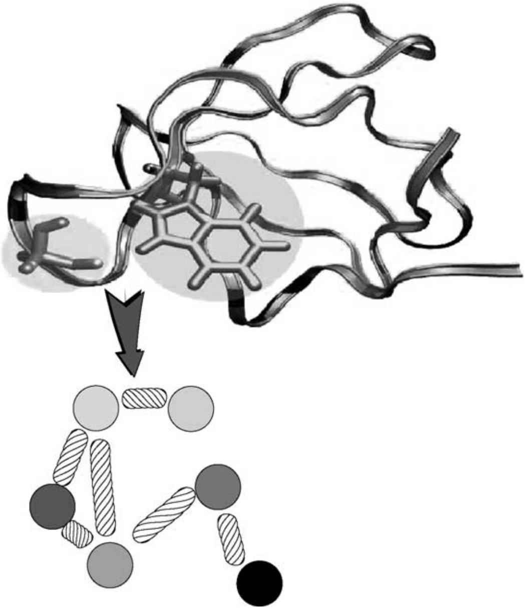

FIG. 5. Constructing protein graphs from protein conformations. Each node corresponds to an amino acid. We draw an edge between any two nodes of a graph if there exists a contact between amino acids, corresponding to these nodes. The contact between two amino acids is defined by the spatial proximity of β carbons (C_α for Gly) of these amino acids. The contact distance is taken to be 8.5 Å.

To answer this question, we hypothesize that[120] the actual topological properties of pre- and posttransition conformations are different. To test this hypothesis, we construct a protein graph (Fig. 5), the nodes of which represent amino acids and the edges of which represent pairs of amino acids that are within the contact range from each other. For SH3, we define the maximum distance between C_β atoms at which a contact exists at 7.5 Å.[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 in Vendruscolo *et al.*[150] and later used for discriminating pre- and posttransition states of SH3 and CI2 proteins:

[149] R. L. Jernigan and I. Bahar, *Curr. Opin. Struct. Biol.* **6,** 195 (1996).

[150] M. Vendruscolo, N. V. Dokholyan, E. Paci, and M. Karplus, *Phys. Rev. E* **65,** 061910 (2002).

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

where N is the number of amino acids and ℓ_{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 in pre- and post-transition states, supporting the hypothesis of ref.[120] 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 Vendruscolo *et al.*,[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, α-spectrin SH3 domain, CI2, and protein L. Using Monte Carlo simulations of the hydrophobic protein model, Treptow *et al.*[151] also suggested the role of protein topology in folding kinetics.

Conclusion

The revolution in protein crystallography has resulted in the identification of a large number of protein structures. The latter became an invaluable source of information on amino acid interactions. We review extensive studies that uncovered the important role of protein topology in folding kinetics. These studies suggested that one can determine protein-folding kinetics to a reasonably detailed level from the knowledge of crystal structure.

We describe analytical and computational tools for determining and characterizing protein-folding kinetics from crystal structures. These include 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 in the native and unfolded states. We find an important role of the L24–G54 contact for the folding kinetics of SH3 proteins. A possible test of the kinetic importance of the L24–G54 contact may come from cross-linking this contact and understanding whether cross-linking stabilizes the native state of the Src SH3.

The hallmark of the relationship between protein crystal structures and their folding kinetics was signified by the success of the Gō model of amino

[151] W. L. Treptow, M. A. A. Barbosa, L. G. Garcia, and A. F. P. de Araújo, *Proteins* **49,** 167 (2002).

acid interactions to study protein folding. We describe several studies that are based on the Gō model. In one such study, we identify the most evasive protein-folding 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 are related to protein topology and, thus, may explain the kinetics of the folding process. These studies unveil the expanding possibilities for studying protein-folding kinetics from their crystal structures.

Acknowledgments

We gratefully acknowledge E. Deeds and A. F. P. de Araújo for critical reading of the manuscript. This work is supported by the NSF and Petroleum Research Fund (to H.E.S.), and by the NIH (GM52126 to E.S.). N.V.D. is supported by an NIH NRSA Grant (GM20251–01).