Kinetics of the protein folding transition

Nikolay V. Dokholyan[a], Sergey V. Buldyrev[b], H. Eugene Stanley[b], and Eugene I. Shakhnovich[a]

[a] *Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA*

[b] *Center for Polymer Studies, Physics Department, Boston University, Boston, MA 02215, USA*

AIP Proceedings of the Third Tohwa University Conference on Statistical Physics, November 8–12, 1999

Abstract. The ability of proteins to find their unique three-dimensional structure in a millisecond to a second time scale has been one of the intriguing problems of biophysics. This problem, pointed out by Levinthal, is known as the protein folding problem. We discuss a nucleation scenario for protein folding that has been proposed to resolve the Levinthal paradox.

I INTRODUCTION

Proteins [1–3] are among the most important building blocks of life. They are responsible for many functions in cell organization, reproduction, signal transduction, and cell death (apoptosis). Proteins carry out transport and storage in living cells and inhibit or catalize chemical reactions. The most intriguing fact about proteins is that their functions in living cells are determined by their three-dimensional structure.

Proteins are linear heteropolymers built by a combination of twenty amino acids (residues), joined by peptide bonds (see Fig. 1). The linear polymer formed by amino acid residues linked by peptide bonds, $H-(NH-C_\alpha HR_i-CO-)_nOH$ is also called a *polypeptide chain* (R_i denotes an amino acid). The number of amino acids in proteins ranges from 40 to 3000.

Only the primary structure (sequence of amino acids) of proteins is encoded by DNA, i. e. only the primary structure of proteins is genetically transfered by DNA. Thus, proteins must themselves find their unique native configuration from the primary structure. However, if one takes a small protein of 100 residues and assumes that each residue can be positioned, on average, in 6 different ways relative to its

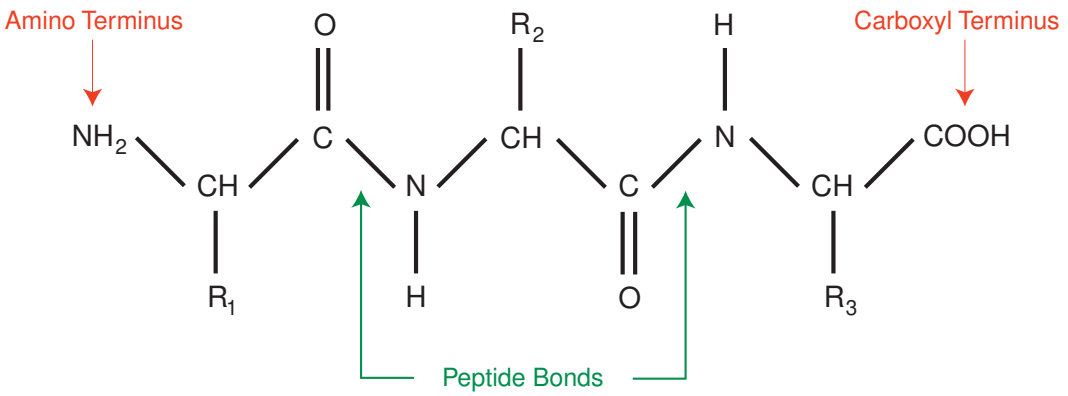

FIGURE 1. Schematic representation of the polypeptide chain. R_1, R_2, etc. are the side groups attached to α-carbons (C_α) of the amino acids.

chain neighbors[1], then the number of possible three-dimensional conformations of such a protein will be $6^{100} \approx 6 \cdot 10^{77}$. The vibrational mode of proteins is of the order of picoseconds, so even if one takes 1 ps per conformation of a residue, the folding by random search will take $6 \cdot 10^{64}$ s, which is roughly $2 \cdot 10^{55}$ years. However, the folding time is of the order of milliseconds to seconds. This paradox was first described by Levinthal in 1968 [4]. Therefore, there should be some conformational "information" stored in a sequence of amino acids, which drives proteins to their native conformation. This information is a superposition of quantum mechanical properties of amino acids.

II THE CONCEPT OF NUCLEATION

Understanding the relevance of the interactions between amino acids to the folding process of a protein is a complex task that has been the subject of a number of theoretical and experimental studies in the past few decades [2, 5–19]. Many studies have been dedicated to identification of the dominant interactions between amino acids, i. e. forces that drive a protein to the native state (see e. g. Refs. [14, 20–22]).

The transition of short proteins from the unfolded to the folded state is accompanied by a drastic reduction of the entropy. It was proposed that the folding transition for short proteins is analogous to the nucleation process at a first order transition [6, 7, 10, 11, 19]. In this scenario, there is competition between two minima of the free energy, the folded state with low energy and entropy and the unfolded state corresponding to high energy and entropy. Specific scenarios, in which a fraction of the contacts play an important role in reducing the time scales for folding, have been proposed by different authors [6, 7, 10, 11] and confirmed experimentally

[1)] This number is surely an underestimation. It is chosen just to make a lower-bound estimation for the "random search" folding time.

[23–25]. Thus, the determination of this set of contacts is a step forward towards the solution of the direct protein folding problem.

Thermodynamically, the folding transition in small proteins is analogous to a first-order transition whereby two thermodynamic states [26] (folded and unfolded) are free energy minima while intermediate states are unstable. The kinetic mechanism of transitions from the unfolded state to the folded state is nucleation [11, 27–30].

Definition: *Folding nuclei* can be defined as the minimal stable element of structure whose existence results in subsequent rapid assembly of the native state.

This definition corresponds to a "post-critical nucleus" related to the first stable structures that appear immediately after the transition state is overcome [31]. The thermal probability of a transition state conformation is low compared to the folded and unfolded states, which are both accessible at the folding transition temperature T_f (see Fig. 2).

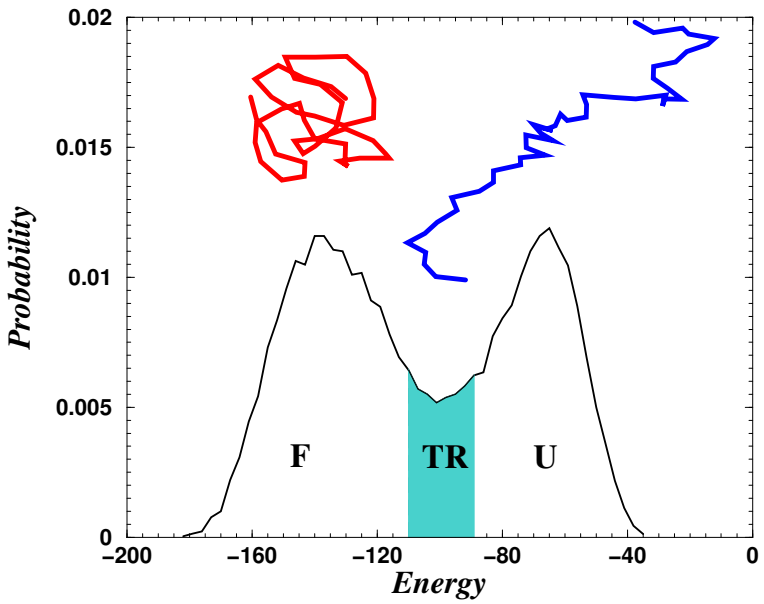

FIGURE 2. The probability distribution of the energy states E of the 46-mer [19] maintained at the folding transition temperature $T_f = 1.44$. The bimodal distribution indicates the presence of two dominant states: the folded (region F) and the unfolded (region U) states. The transition state ensemble belongs to region TR of the histogram $\{-110 < E < -90\}$. The insets show typical conformations in the folded and unfolded regions.

Kinetic analyses [6, 11, 30–34] for a number of lattice model chains of different lengths and degrees of sequence design (optimization) point to a specific protein folding nucleus scenario. Passing through the transition state with subsequent

rapid assembly of the native conformation requires the formation of some (small) number of specific obligatory contacts (protein folding nucleus). This result has been verified [6] for sequences designed in the lattice model using different sets of potentials, where it is shown that nucleus location was identical for two different sequences designed with different potentials to fold into the same structure of a lattice 48-mer. This finding and related results [30, 35] suggest that the folding nucleus location depends more on the topology of the native structure than on a particular sequence that folds into that structure.

The dominance of geometrical/topological factors in the determination of the folding nucleus is a remarkable property that has evolutionary implications. It is important to understand the physical origin of this property of folding proteins and to assess its generality. To this end, it is important to study other than lattice models and other than Monte-Carlo dynamic algorithms. To search for the nucleus in a continuous off-lattice model, the discrete molecular dynamics (MD) simulation technique [7, 36, 37] (the Gō model [5, 12, 38] with the square-well potential of the inter-residue interaction) is employed.

The proposed in [19] method to search for a folding nucleus is based on the observation [31] that equilibrium fluctuations around the native conformation can be separated into "local" unfolding (followed by immediate refolding) and "global" unfolding that lead to a transition into an unfolded state and requires more time to refold. Local unfolding fluctuations are the ones that do not reach the top of the free energy barrier and, hence, are committed to moving quickly back to the native state. In contrast, global unfolding fluctuations are the ones that overcome the barrier and are committed to descend further to the unfolded state. Similarly, the fluctuations from the unfolded state can be separated into those that descend back to the unfolded state and those that result in productive folding. The difference between the two modes of fluctuation is whether or not the major free energy barrier is overcome. This means that the nucleation contacts (i. e. the ones that are formed on the "top" of the free energy barrier as the chain passes it upon folding) should be identified as contacts that are present in the "maximally locally unfolded" conformations but are lost in the globally unfolded conformations of comparable energy.

In order to identify the folding nucleus, in Ref. [19] we studied the conformations of the 46-mer that appear in various kinds of folding $\rightleftharpoons$ unfolding fluctuations. First, consider the time behavior of the potential energy at T_f. The transition state conformations belong to the transition region TR from the folded state to the unfolded state that lies in the energy range between states corresponding to the folded and unfolded conformations (see Fig. 2). Region TR corresponds to the minimum of the histogram of the energy distribution. If one knows the past and the future of a certain conformation that belongs to the TR, then there are 2 distinct types of such conformations: UU conformations that originate in and return to the unfolded region without ascending to the folded region; and FF conformations that originate in and return to the folded region without descending to the unfolded region.

It was argued [19] that if the nucleus exists, then the FF, and UU conformations must have different properties depending on their history. A difference between the FF conformations and UU conformations is that the protein folding nucleus is more likely to be retained in the FF conformations than in the UU conformations. Correspondingly, the difference between the frequencies of all native contacts in FF versus UU events has been determined in Ref. [19] and it was found that there is a set of 5 contacts for which this difference is most pronounced. Additional tests in Ref. [19] for the selected set of nucleic contacts presented reasonable evidence for the nucleation scenario.

III DISCUSSION

The main conclusion is that the existence of a few specific contacts is a signature of the transition state conformations. Those contacts can be defined as the protein folding nucleus. Other contacts may also be present in transition state conformations. However, they are optional and vary from conformation to conformation, while nucleation contacts are present in transition state conformations with high probability. Formation of nucleation contacts can be considered an obligatory step in the folding process: after they are formed the major barrier is overcome and subsequent folding proceeds "downhill" in the free energy landscape without encountering any further major free energy barriers.

The protein folding nucleus scenario of the transition state was initially derived from Monte-Carlo studies of lattice models [6, 30, 31] and was consistent with protein engineering experiments with several small proteins [23–25, 39]. In [19], for the first time, this scenario is confirmed in off-lattice MD simulations. The consistency between conclusions made in different simulations [6, 30, 31] and in experiments [23–25, 39] is remarkable, and supports the possibility that the protein folding nucleus formation is a generic scenario to describe the protein folding transition state.

IV CONCLUSION

The knowledge of the thermodynamics and kinetics of the protein folding process and basic mechanisms that govern the folding process might aid us to solve the protein folding problem. The nucleation scenario that was found plausible in off-lattice MD simulations of Ref. [19], is a step towards solution of this problem. The ability to understand the folding properties of proteins will not only allow us to understand protein function and develop new drugs, but also to open the door to yet unforeseen discoveries.

REFERENCES

1. T. Creighton, *Proteins: structures and molecular properties*, W. H. Freeman and Co., New York, ii edition, 1993
2. M. Karplus and E. I. Shakhnovich. Protein folding: theoretical studies of thermodynamics and dynamics. in T. Creighton, ed., *Protein Folding*. W. H. Freeman and Company, New York 1994.
3. O. B. Ptitsyn. The molten globule state. in T. Creighton, ed., *Protein Folding*. W. H. Freeman and Company, New York 1994.
4. C. Levinthal, Are there pathways for protein folding?, *J. Chim. Phys.* **65**, 44 (1968)
5. N. Gō and H. Abe, Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation, *Biopolymers* **20**, 991–1011 (1981)
6. E. I. Shakhnovich, V. I. Abkevich and O. Ptitsyn, Conserved residues and the mechanism of protein folding, *Nature* **379**, 96–98 (1996)
7. N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley and E. I. Shakhnovich, Molecular dynamics studies of folding of a protein-like model, *Folding & Design* **3**, 577–587 (1998)
8. J. D. Bryngelson and P. G. Wolynes, Intermediates and barrier crossing in a random energy model (with applications to protein folding), *J. Phys. Chem.* **93**, 6902–6915 (1989)
9. C. B. Anifsen, Principles that govern the folding of the protein chains, *Science* **181**, 223–230 (1973)
10. N. Gō, Theoretical studies of protein folding, *Ann. Rev. Biophys. Bioeng.* **12**, 183–210 (1983)
11. E. I. Shakhnovich, Theoretical studies of protein-folding thermodynamics and kinetics, *Curr. Opinion Struc. Biol.* **7**, 29–40 (1997)
12. H. Taketomi, Y. Ueda and N. Gō , Studies on protein folding, unfolding and fluctuations by computer simulations, *Int. J. Peptide Protein Res.* **7**, 445 (1975)
13. K. A. Dill, Theory for the folding and stability of globular proteins, *Biochemistry* **24**, 1501–1509 (1985)
14. H. Li, C. Tang and N. S. Wingreen, Nature of driving force for protein folding: A result from analyzing the statistical potential, *Phys. Rev. Lett.* **79**, 765–768 (1997)
15. D. K. Klimov and D. Thirumalai, Criterion that determines the foldability of proteins, *Phys. Rev. Lett.* **76**, 4070–4073 (1996)
16. D. Hoffmann and E.-W. Knapp, Polypeptide folding with off-lattice Monte Carlo dynamics: the method, *Eur. Biophys. J.* **24**, 387–403 (1996)
17. M. Vásquez, Modeling side-chain conformation, *Curr. Opinion Struc. Biol.* **6**, 217–221 (1996)
18. C. Micheletti, J. R. Banavar, A. Maritan and F. Seno, Protein structures and optimal folding emerging from a geometrical variational principle, *prl* **82**, 3372–3375 (1998)
19. N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley and E. I. Shakhnovich, Identifying the protein folding nucleus using molecular dynamics, *J. Mol. Biol.* **296**, 1183–1188 (2000)
20. S. Miyazawa and R. L. Jernigan, Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation, *Macromolecules* **18**,

534–552 (1985)
21. K. A. Dill, Dominant forces in protein folding, *Biochemistry* **29**, 7133–7155 (1990)
22. R. L. Jernigan and I. Bahar, Structure-derived potentials and protein simulations, *Curr. Opinion Struc. Biol.* **6**, 195–209 (1996)
23. J. C. Martinez, M. T. Pissabarro and L. Serrano, Obligatory steps in protein folding and the conformational diversity of the transition state, *Nature Struct. Biol.* **5**, 721–729 (1998)
24. J. L. Neira, P. Sevilla, M. Menendez, M. Bruix and M. Rico, Hydrogen exchange in ribonuclease A and ribonuclease S: Evidence for residual structure in the unfolded state under native conditions, *J. Mol. Biol.* **285**, 627–643 (1999)
25. D. E. Callihan and T. M. Logan, Conformations of peptide fragments from the FK506 binding protein: Comparison with the native and urea-unfolded states, *J. Mol. Biol.* **285**, 2161–2175 (1999)
26. S. E. Jackson, How do small single-domain proteins fold?, *Folding & Design* **3**, R81–R91 (1998)
27. V. G. Karpov and D. W. Oxtoby, Nucleation in disordered systems, *Phys. Rev. B* **54**, 9734–9745 (1996)
28. E. M. Lifshitz and L. P. Pitaevskii, *Physical kinetics*, Pergamon Press, Oxford, 1981
29. A. R. Fersht, Nucleation mechanisms in protein folding, *Curr. Opinion Struc. Biol.* **7**, 3–9 (1997)
30. V. S. Pande, A. Yu. Grosberg, D. S. Rokshar and T. Tanaka, Pathways for protein folding: is a new view needed?, *Curr. Opinion Struc. Biol.* **8**, 68–79 (1998)
31. V. I. Abkevich, A. M. Gutin and E. I. Shakhnovich, Specific nucleus as the transition state for protein folding: evidence from the lattice model, *Biochemistry* **33**, 10026–10036 (1994)
32. E. I. Shakhnovich, Folding nucleus: specific of multiple? Insights from lattice models and experiments, *Folding & Design* **3**, R108–R111 (1998)
33. L. A. Mirny, V. I. Abkevich and E. I. Shakhnovich, How evolution makes proteins fold quickly, *Proc. Natl. Acad. Sci. U. S. A.* **95**, 4976–4981 (1998)
34. A. M. Gutin, V. I. Abkevich and E. I. Shakhnovich, A protein engineering analysis of the transition state for protein folding: simulation in the lattice model, *Folding & Design* **3**, 183–194 (1998)
35. D. K. Klimov and D. Thirumalai, Lattice models for proteins reveal multiple folding nuclei for nucleation-collapse mechanism, *J. Mol. Biol.* **282**, 459–480 (1998)
36. Y. Zhou, M. Karplus, J. M. Wichert and C. K. Hall, 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)
37. Y. Zhou and M. Karplus, Folding thermodynamics of a three-helix-bundle protein, *Proc. Natl. Acad. Sci. U. S. A.* **94**, 14429–14432 (1997)
38. H. Abe and N. Gō, Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins, *Biopolymers* **20**, 1013–1031 (1981)
39. L. S. Itzhaki, D. E. Otzen and A. R. Fersht, 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)