The protein folding problem

Nikolay V. Dokholyan*
Department of Chemistry & Chemical Biology, Harvard University, 12 Oxford Street, Cambridge, MA 02138, USA

(Recent Res. Devel. Stat. Phys. **1:** 77–84 (2000))

We present an overview of the protein folding problem. We discuss difficulties associated with this problem, its importance, the various approaches to its solution, and recent advances in the field. We discuss two important aspects of protein folding properties: its folding thermodynamics and kinetics. We also describe a popular protein folding scenario – folding via nucleation.

I. INTRODUCTION

Proteins [1–3] are among the most important building blocks of life (Fig. 1). 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. Proteins 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. 2). 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, 6 different ways relative to its 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 miliseconds to seconds. This paradox was first described by Levinthal in 1968 [4]. Therefore, there must be some conformational "information" stored in a sequence of amino acids, which drives proteins to their native conformation. This information is a superposition of the quantum mechanical properties of amino acids.

II. PROTEIN FOLDING PROBLEM

A simplified view of proteins is that they are molecules with specific three-dimensional structures of specific functionalities that are determined by the one-dimensional sequences of amino acids. There are two principal questions that arise from this view: *(i) The "direct" folding problem:* How can one predict the three-dimensional conformation of a protein based on its sequence of amino acids? *(ii) The "inverse" or design folding problem:* How can one predict a sequence of amino acids based on the three dimensional structure of the protein?

A. Inverse (design) protein folding problem

Experimental approaches to protein design have had only limited success, providing polypetides that could fold into compact but mostly disordered conformations [5,6]. The bottleneck in the protein design problem is that the number of sequences grows exponentially with the sequence length of the protein. For example, there are $20^{100} \approx 10^{130}$ possible 100 amino acid sequences. The fraction of sequences that fold into protein-like structures is negligible. Thus, the probability that one will select, at random, a sequence that will fold into a protein-like structure is vanishingly

**Correspondence to:* dokh@wild.harvard.edu

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

low. Convincing success in the protein design problem may come from reliable theoretical approaches that make it possible to find a sequence that folds to a unique stable native structure.

It has been pointed out [6] that *experimental protein design* is limited due to the poor corroboration of theory and experiment. Recent success of Dahiyat and Mayo [7] in the design of a small protein is based on the synergism of theory and experiment. This work demonstrates the importance of theoretical approaches in protein design. The limitation of this work comes from the protein sequence length. Their approach requires complete enumeration of all sequence candidates, which becomes exponentially difficult as sequence length increases. For example, Dahiyat and Mayo [7] used a library of sequences of the size 1.9×10^{27} for the target 28 amino acid $\beta\beta\alpha$-motif, structurally similar to the second zinc finger module of the DNA binding protein Zif268. The success of Dahiyat and Mayo [7] calls for further refinements of the theoretical approaches to the design problem.

Because the number of possible sequences is enormous and the fraction of sequences which fold into a given structure [8] is negligibly small, the probability that one will find a sequence at random is also negligible. Even clever experimental approaches, such as *phage display* [9], which bias experimental sequence searches towards targets, suffer from technical limitations on the total number of sequences that can be surveyed.

Several design algorithms have been proposed (see e. g. [10,11] and also review [6]). A Monte Carlo algorithm [12,13] seems to be the most advantageous [6] among all other algorithms since it converges to the canonical distribution and its results can be understood from the statistical mechanics viewpoint [14–17]. Thus, the tools of statistical mechanics become relevant to protein design.

B. Direct protein folding problem

The three-dimensional structure of globular proteins in cells is determined by their amino acid composition. 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 [2,18–32]. 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 (NS) (see e. g. Refs. [27,33–35]).

The transition of short proteins from the unfolded to the folded state is accompanied by a drastic reduction of entropy (see Fig. 3 and Fig. 1 in [24]). It was proposed that the folding transition for short proteins is analogous to the nucleation process at a first order transition [19,20,23,24,32]. 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. Levinthal pointed out that the time scale for a random search in the space of protein configurations would be incompatible with biological folding times [4]. 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 [19,20,23,24] and confirmed experimentally [36–38]. Thus, the determination of this set of contacts is a step towards the solution of the direct protein folding problem.

C. Why it is hard to solve the protein folding problem

The vast dimensionality of the protein conformational space [4] makes the folding time too long to be reachable by direct computational approaches [2,23,24]. While the folding of a real protein occurs on the time scale of 1ms - 1s, the traditional molecular dynamics (MD) algorithms are able to resolve only the nanoscale time region. More realistic MD, based on the quantum mechanical calculations of the states of all atoms, limit the study only to polypeptides of five amino acids in length. However, the stability condition requires that the chain contains at least 20 amino acids, since the stabilization energy is of the order of 0.1 k_BT per amino acid [34], where k_BT is the average energy of the environment.

Simplified models [14,18,19,21,25,26,34,39–41] became popular due to their ability to reach reasonable time scales and to reproduce the basic thermodynamic and kinetic properties of real proteins [1,2,42]: *(i)* unique native state, i. e. there should exist a single conformation with the lowest potential energy; *(ii)* cooperative folding transition (resembling first order transition); *(iii)* thermodynamical stability of the native state; *(iv)* kinetic accessibility, i. e. the native state should be reachable in a biologically reasonable time [28,41].

Monte Carlo (MC) simulations on lattices (see, e. g., [18,24,25,39] and references therein) appear to be useful for studying theoretical aspects of protein folding. The Monte Carlo algorithm is based on a set of rules for the transition from one conformation to another. These transitions are weighted by some transition matrix, which reflects the phenomena under study. The simplicity of the algorithm and the significantly small conformational space of the protein models (due to the lattice constraints) make MC on-lattice simulations a powerful tool for studying the equilibrium dynamics of protein models.

However, lattice models impose strong constraints on the angles between the covalent bonds, thereby greatly restricting the conformational space of the protein-like model. The additional drawback of this restriction lies in the poor capability of these models to discern the geometrical properties of the proteins.

Time in MC algorithms is estimated as the average number of moves (over an ensemble of the folding $\rightleftharpoons$ unfolding transitions) made by a model protein. It was pointed out [43] that MC simulations are equivalent to the solution of the master equation for the dynamics, so there is a relationship between physical time and computer time, which is counted as the number of MC steps. However, a number of delicate issues — such as the dependence of the dynamics on the set of allowed MC moves — remain outstanding, so an independent test of the dynamics using the MD approach is needed.

To address the questions with sensitivity to geometrical details, it is useful to study off-lattice models of protein folding. Thus far, several off-lattice simulations have been performed [44–46], which demonstrate the ability of the simplified models to study protein folding.

D. A rapid tool to perform off-lattice simulations

Recently, a new approach for simulations of model proteins, discrete molecular dynamics (DMD), has been implemented [20,32,47,48] to study the dynamics of homo- and heteropolymers (proteins). This approach permits the rapid testing of the folding properties of proteins within reasonable processor time. This MD algorithm has proven to be a powerful tool to study the thermodynamics and kinetics of the folding $\rightleftharpoons$ unfolding transition of simplified models of proteins [18,39].

The potential of interaction between pairs of residues is modeled by a “square-well”, which allows an increase in the speed of the simulations [20,32,47,48]. This approach is based on the tracing of collision events in the simulated system, such that it is not necessary to follow the dynamics of the system during times when there are no events (i.e., collisions). This is a crucial difference between traditional MD and DMD. In addition to increasing the speed of the simulation, the DMD algorithm allows for the tracking of “realistic” (not discretized) time. It has been shown that the DMD algorithm is able to resolve in time the collapse transition for a wide range of temperatures.

In addition, such an algorithm can be a useful compromise between computationally heavy traditional MD and fast, but restrictive MC. It has been demonstrated in [20] that a model protein reproduces the principal features of folding phenomena *(i)* – *(iv)* described in Section II C.

E. A path to solution of the protein folding problem

There are two questions that might shed light on the protein folding problem and are the subject of this dissertation.

1. What are the thermodynamic properties of proteins and what are the structural properties of the protein near the folding transition? For example, it has been shown [20] that more than 50% of all of the inter-residue contacts are present near the folding transition temperature T_f. This set of contacts is called the *core* of the protein.

   Definition 1 The *core* is a subset of residues, which maintains the backbone of the structure at temperatures close to the folding transition temperature.

   Identification of the protein core is a challenging task, both in real experiments and computer simulations. A method describing the identification of the core of a model 65-mer has been described in [20].

2. What are the kinetic properties of proteins at the folding transition? This question is specifically important since it is directly connected to the mechanism of protein folding. Several scenarios of protein folding have been proposed [6,23,32,40,49–53]. Simulations of the off-lattice model 46-mer [32] confirmed the nucleation scenario.

   Definition 2 The *nucleus* is the structure (few contacts $< 5\%$) that appears in the transition state and results in rapid folding. If the nucleus is disrupted in the native state, then the structure is unfolded.

The difference between the core and the nucleus of a protein is crucial: while the *core* is a persistent part of the structure at equilibrium, the *nucleus* is a fragment of the protein structure, which is assembled in the transition state (TS) — the folding $\rightleftharpoons$ unfolding barrier (see Fig. 3 and Fig. 1 in [24]).

In addition, the nucleus is defined kinetically, and has “physical” meaning – like bubbles that appear in liquid during boiling. On the contrary, the core is defined thermodynamically, and refers to the geometrical properties of the protein which persist near the folding transition temperature.

III. THE SIGNIFICANCE OF THE PROTEIN FOLDING PROBLEM

The solution of the direct and inverse protein folding problems will impact modern medicine, biology, and biomolecular engineering [1]. One important clinical outcome would be effective drug design. For example, Benjamin and McMillan [54] underline the crucial significance of the heat shock family of stress proteins (HSPs) in cardiovascular conditions such as cardiac hypertrophy, vascular wall injury, ischemic preconditioning, and aging. HSPs serve as part of the defense strategy to ensure survival of the cell. They are engaged in response to various physiological stresses (e. g., heat, hemodynamics, mutant proteins, and oxidative injury) and in decisions to repair or degrade misfunctional, damaged proteins. Therefore, the development of proteins similar in function to the HSPs may aid the protection of essential cellular proteins.

In addition, solving the protein design problem would shed light on the relation between protein structure and function. For example, the mechanism of muscle contraction rests upon several proteins, namely myosin and actin filaments [55] and titin [56,57]. Though the physical mechanisms of these proteins are thought to be known, the global picture of muscle contraction is still unresolved.

Specific interest has been generated by the special class of transmembrane proteins, *integrins*, responsible for the transduction of the mechanical signals between cell and extracellular matrix (ECM), cell-cell adhesion, and also involved in a vast variety of intercellular phenomena (e. g. *“tensegrity”*) [58]. The distinguished feature of integrins and titins is the strong binding affinity due to the special region called *type III repeat* [59]. This motif constitutes 2% of all proteins in the cell. *Type III repeat* fold will be used as a target model in my design procedure.

Important breakthroughs may also come from the understanding of *(i)* the expression mechanisms of growth factors, e. g. vascular endothelial growth factor (VEGF) [60], which are synthesized and secreted by differentiated cells in response to stimuli including hypoxia; *(ii)* the function of the low density lipoprotein (LDL) receptor [61], which mediates cellular uptake of lipoprotein particles. The LDL receptor plays a key role in cholesterol homeostasis.

The role of proteins in the cell is essential. Proteins are involved in most cellular and intercellular functions. The above mentioned proteins constitute just a small subset of all the proteins involved in various structural, regulatory, and defense processes. The knowledge of the thermodynamics and kinetics of the protein folding process and the basic mechanisms that govern the folding process might aid us to solve the protein folding problem. The ability to predict the structure of a protein from its sequence and the ability to predict the sequence of a protein from its structure will not only allow us to understand protein function and develop new drugs, but also to open the door to yet unforeseen discoveries.

We thank R. S. Dokholyan for careful reading of the manuscript. NVD is supported by NIH postdoctoral fellowship (GM20251-01).

[1] Creighton, T., *Proteins: structures and molecular properties*, W. H. Freeman and Co., New York, ii edition, 1993

[2] Karplus, M. and 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.

[3] Ptitsyn, O. B. The molten globule state. in Creighton, T., ed., *Protein Folding.* W. H. Freeman and Company, New York 1994.

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

[5] Quinn, T., Tweedy, N., Williams, R., Richardson, R. and Richardson, D., *Proc. Natl. Acad. Sci. U. S. A.* **91**, 8747–8751 (1994)

[6] Shakhnovich, E. I., *Folding & Design* **3**, R45–R58 (1998)

[7] Dahiyat, B. I. and Mayo, L., *Science* **278**, 82–87 (1997)

[8] Finkelstein, A. V., Gutin, A. and Badretdinov, A., *Proteins* **23**, 142–149 (1995)

[9] Riddle, D. S., Santiago, J. V., Bray-Hall, S. T., Doshi, N., Grantcharova, V. P., Yi, Q. and Baker, D., *Nature Struct. Biol.* **4**, 805–809 (1997)

[10] Jones, D. T., *Curr. Opinion Biotechnol.* **6**, 452–459 (1995)

[11] Koehl, P. and Delarue, M., *Curr. Opinion Struc. Biol.* **6**, 222–226 (1996)

[12] Allen, M. P. and Tildesley, D. J., *Computer simulation of liquids*, Clarendon Press, Oxford, 1987

[13] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N, Teller, A. and Teller, E., *J. Chem. Phys.* **21**, 1087–1092 (1953)

[14] Shakhnovich, E. I., *Phys. Rev. Lett.* **72**, 3907–3910 (1994)

[15] Shakhnovich, E. I. and Gutin, A. M., *Proc. Natl. Acad. Sci. U. S. A.* **90**, 7195–7199 (1993)

[16] Shakhnovich, E. I. and Gutin, A. M., *Protein Eng.* **6**, 793–800 (1993)

[17] Saven, J. and Wolynes, P., *J. Chem. Phys.* **101**, 8375–8389 (1997)

[18] Gō, N. and Abe, H., *Biopolymers* **20**, 991–1011 (1981)

[19] Shakhnovich, E. I., Abkevich, V. I. and Ptitsyn, O., *Nature* **379**, 96–98 (1996)

[20] Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. and Shakhnovich, E. I., *Folding & Design* **3**, 577–587 (1998)

[21] Bryngelson, J. D. and Wolynes, P. G., *J. Phys. Chem.* **93**, 6902–6915 (1989)

[22] Anifsen, C. B., *Science* **181**, 223–230 (1973)

[23] Gō, N., *Ann. Rev. Biophys. Bioeng.* **12**, 183–210 (1983)

[24] Shakhnovich, E. I., *Curr. Opinion Struc. Biol.* **7**, 29–40 (1997)

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

[26] Dill, K. A., *Biochemistry* **24**, 1501–1509 (1985)

[27] Li, H., Tang, C. and Wingreen, N. S., *Phys. Rev. Lett.* **79**, 765–768 (1997)

[28] Klimov, D. K. and Thirumalai, D., *Phys. Rev. Lett.* **76**, 4070–4073 (1996)

[29] Hoffmann, D. and Knapp, E.-W., *Eur. Biophys. J.* **24**, 387–403 (1996)

[30] Vásquez, M., *Curr. Opinion Struc. Biol.* **6**, 217–221 (1996)

[31] Micheletti, C., Banavar, J. R., Maritan, A. and Seno, F., *Phys. Rev. Lett.* **82**, 3372–3375 (1998)

[32] Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. and Shakhnovich, E. I., *J. Mol. Biol.* **296**, 1183–1188 (2000)

[33] Miyazawa, S. and Jernigan, R. L., *Macromolecules* **18**, 534–552 (1985)

[34] Dill, K. A., *Biochemistry* **29**, 7133–7155 (1990)

[35] Jernigan, R. L. and Bahar, I., *Curr. Opinion Struc. Biol.* **6**, 195–209 (1996)

[36] Martinez, J. C., Pissabarro, M. T. and Serrano, L., *Nature Struct. Biol.* **5**, 721–729 (1998)

[37] Neira, J. L., Sevilla, P., Menendez, M., Bruix, M. and Rico, M., *J. Mol. Biol.* **285**, 627–643 (1999)

[38] Callihan, D. E. and Logan, T. M., *J. Mol. Biol.* **285**, 2161–2175 (1999)

[39] Abe, H. and Gō, N., *Biopolymers* **20**, 1013–1031 (1981)

[40] Abkevich, V. I., Gutin, A. M. and Shakhnovich, E. I., *Biochemistry* **33**, 10026–10036 (1994)

[41] Gutin, A. M., Abkevich, V. I. and Shakhnovich, E. I., *Proc. Natl. Acad. Sci. U. S. A.* **92**, 1282–1286 (1995)

[42] Privalov, P. L., *Ann. Rev. Biophys. Biophys. Chem.* **18**, 47–69 (1989)

[43] Baumgartner, A., ed., *Applications of the Monte-Carlo simulations in statistical physics*, Springer-Verlag, New York, 1987

[44] Irbäck, A. and Schwarze, H., *J. Phys. A: Math. Gen.* **28**, 2121–2132 (1995)

[45] Berriz, G. F., Gutin, A. M. and Shakhnovich, E. I., *J. Chem. Phys.* **106**, 9276–9285 (1997)

[46] Guo, Z. and Brooks, III, C. L., *Biopolymers* **42**, 745–757 (1997)

[47] Zhou, Y., Karplus, M., Wichert, J. M. and Hall, C. K., *J. Chem. Phys.* **107**, 10691–10708 (1997)

[48] Zhou, Y. and Karplus, M., *Proc. Natl. Acad. Sci. U. S. A.* **94**, 14429–14432 (1997)

[49] Karplus, M. and Weaver, D. L., *Nature* **260**, 404–406 (1979)

[50] Scheraga, H. A., *Biopolymers* **20**, 1877–1899 (1981)

[51] Wetlaufer, D. B., *Proc. Natl. Acad. Sci. U. S. A.* **70**, 691–701 (1973)

[52] Wolynes, P., *Folding & Design* **3**, R107 (1998)

[53] Thirumalai, D. and Klimov, D. K., *Folding & Design* **3**, R112–R118 (1998)

[54] Benjamin, I. J. and McMillan, D. R., *Circ. Res.* **83**, 117–132 (1998)

[55] Mendelson, R. and Morris, E. P., *Proc. Natl. Acad. Sci. U. S. A.* **94**, 8533–8538 (1997)

[56] Tskhovrebova, L., Trinick, J., Sleep, J. A. and Simmons, R. M., *Nature* **387**, 308–312 (1997)

[57] Kellermayer, M. S. Z., Smith, S. B., Granzier, H. L. and Bustamante, C., *Science* **276**, 1112–1116 (1997)

[58] Chicurel, M. E., Chen, C. S. and Ingber, D. E., *Curr. Opin. Cell Biol.* **10**, 232–239 (1998)

[59] Dickinson, C. D., Veerapandian, B., Dai, X.-P., Hamlin, R. C., h. Xuong, N., Ruoslahti, E. and Ely, K. R., *J. Mol. Biol.* **236**, 1079–1092 (1994)

[60] Keyt, B. A., Berleau, L. T., Nguyen, H. V., Chen, H., Heinsohn, H., Vandlen, R. and Ferrara, N., *J. Mol. Biol.* **271**, 7788–7795 (1996)

[61] Simmons, T., Newhouse, Y. M., Arnold, K. S., Innerarity, T. L. and Weisgraber, K. H., *J. Biol. Chem.* **272**, 25531–25536 (1996)

[62] Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Jr., E. F. Meyer, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M., *J. Mol. Biol.* **112**, 535–542 (1977)

[63] Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. and Weng, J. Protein Data Bank. in Allen, F. H., Bergerhoff,

G. and Sievers, R., eds., *Crystallographic Databases-Information Content, Software Systems, Scientific Applications*, pp. 107–132. Data Commission of the International Union of Crystallography, Cambridge 1987.

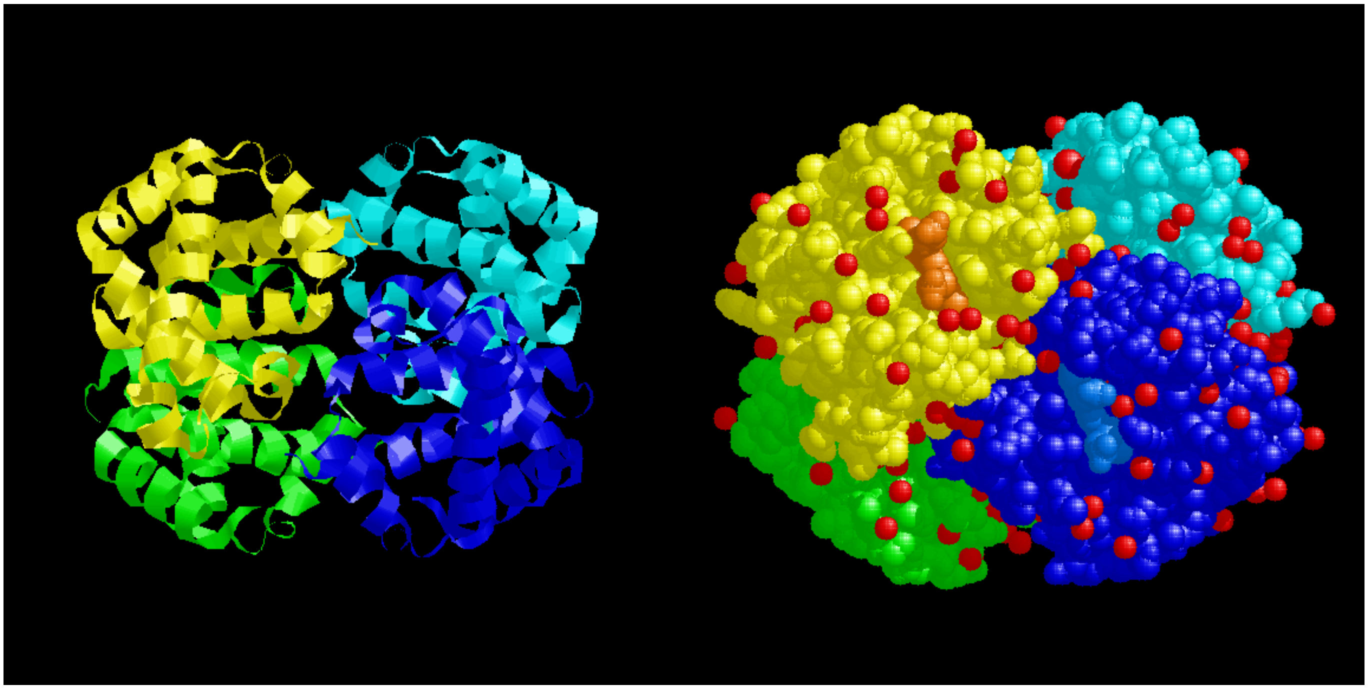

FIG. 1. Hemoglobin (deoxy) molecule (Brookhaven Protein Data Bank [62,63] accession code $2HHB$) is responsible for oxygen transport. The left image is the ribbon representation; the right image is the spacefilling representation.

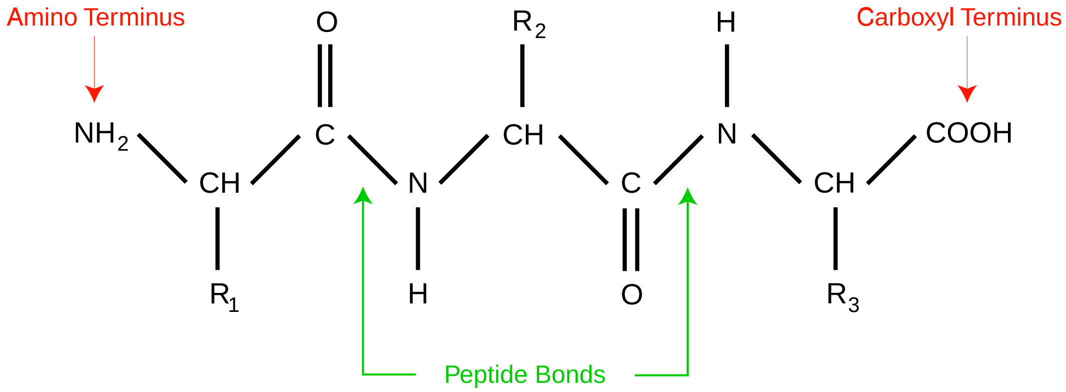

FIG. 2. Schematic representation of the polypeptide chain. R_1, R_2, etc. are the side groups attached to α-carbons (C_α) of the amino acids.

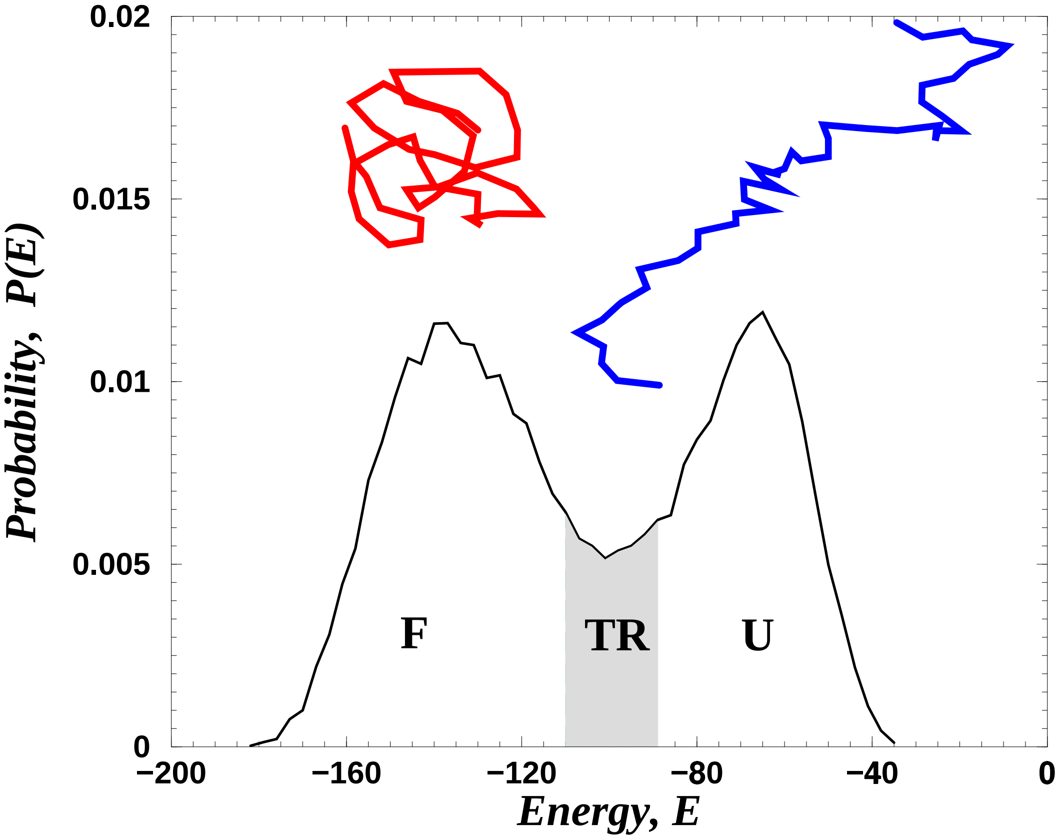

FIG. 3. The probability distribution, $P(E)$, of the energy states E of the 46-mer maintained at the folding transition temperature $T_f = 1.44$ [32]. The bimodal distribution indicates the presence of two dominant states: the folded (region F) and the unfolded (region U) states. The transition state ensemble belongs to region TR of the histogram $\{-110 < E < -90\}$. The insets show typical conformations in the folded and unfolded regions.