Protein Folding

I. Characteristics of proteins

1. Proteins are one of the most important molecules of life. They perform numerous functions, from storing oxygen in tissues or transporting it in a blood (proteins myoglobin and hemoglobin) to muscle contraction and relaxation (titin) or cell mobility (fibronectin) to name a few.

2. Proteins are heteropolymers and consist of 20 different monomers or amino acids (Fig. 1). When amino acids are linked in a chain, they become residues and form a polypeptide sequence. Amino acids differ with respect to their side chains.

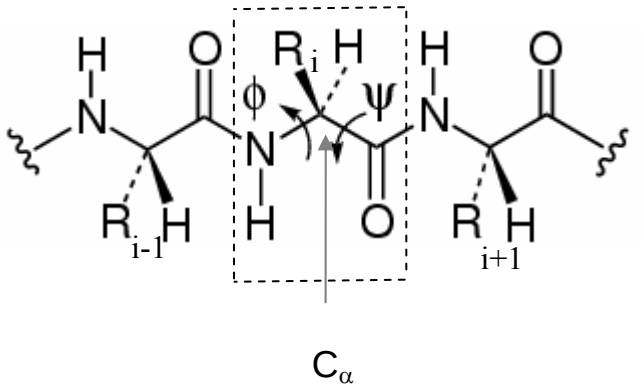


Fig. 1 Structure of a polypeptide chain. A single amino acid residue *i* boxed by a dashed lines contains amide group NH, carboxyl group CO and C_α-carbon. Amino acid side chain R_i along with hydrogen H are attached to C_α-carbon. Amino acids differ with respect to their side chains R_i. Protein sequences utilize twenty different residues.

Roughly speaking, depending on the nature of their side chains amino acids may be divided into three classes – hydrophobic, hydrophilic (polar), and charged (i.e., carrying positive or negative net charge) amino acids. The average number of amino acids N in protein is about 450, but N may range from about 30 to 10^4.

3. The remarkable feature of proteins is an existence of unique native state - a single well-defined structure, in which a protein performs its biological function (Fig. 2). Native states of proteins arise due to the diversity in amino acids. (Illustration of the uniqueness of the native state using lattice model is given in the class.) Large proteins are often folded into relatively independent units called domains. The number of amino acids in a single domain is rarely in excess of 100.

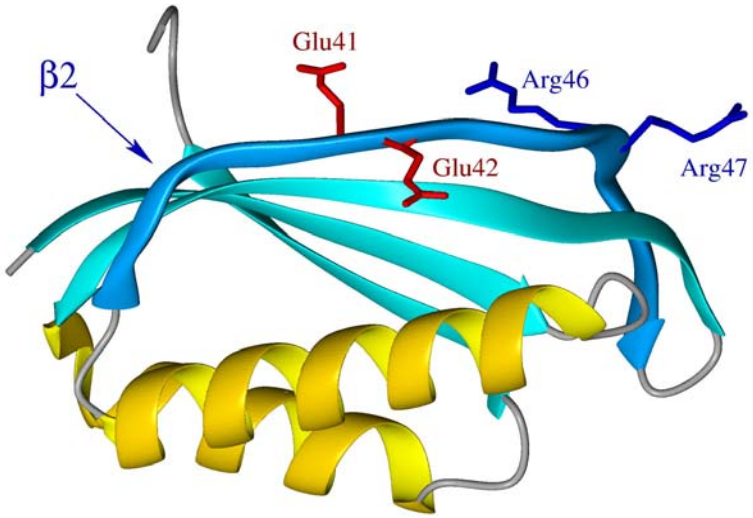


Fig. 2 The native structure of a protein S6 is shown using ribbon representation, which follows the trace of protein sequence in 3D space. Different types of native structure, helix and strands, are color coded in yellow and blue, respectively. Side chains of several charged amino acids are shown in blue (positively charged) and red (negatively charged). S6 contains 101 amino acids and more than 800 heavy atoms.

4. Protein native states are stabilized by electrostatic and van der Waals interactions between protein atoms as well as between proteins atoms and solvent. Complex interplay of these fundamental interactions results in the following effects that determine native structures:

   (i) *steric interactions* are due to excluded volumes of atoms, which prevent two atoms to occupy the same spot in space;

   (ii) *Salt bridges* are electrostatic (Coulombic) interactions between charged amino acids;

   (iii) *Hydrogen bonding* is electrostatic in origin interaction, resulting from sharing a hydrogen between donor (typically, nitrogen) and acceptor (typically, oxygen) atoms.

   (iv) *Hydrophobic interactions* are the outcome of the preferences of hydrophobic residues to avoid water and hydrophilic residues to hydrate. Hydrophobic interactions are effective, which arise solely due to protein-solvent interactions.

Steric interactions are responsible for protein structure on a local scale, such as packing of neighboring side chains against each other. *Hydrogen bonding* largely determines the secondary structure of proteins, such as helices or strands (see S6 picture in Fig. 2). Salt

bridges and hydrophobic interactions maintain tertiary structure of a protein as a whole (tertiary structure encompasses the entire structure of S6 in Fig. 2). It is also important to keep in mind that proteins interact with other biomolecules in a cell (other proteins, DNA, RNA etc). Because cellular interior is very crowded (the occupied volume fraction is about 0.3), these interactions may change, often significantly, the properties of proteins.

Proteins are only marginally stable in their native states. The typical Gibbs free energy difference ΔG between the native and unfolded states is between 2 to 10 *kcal/mol* under normal physiological conditions (300K, normal pH, no denaturants etc) (Recommended reading: S. F. Jackson *Folding & Design* **3**, R81 (1998)).

II. Protein folding problem

Proteins are synthesized on ribosomes by transcription/translation of DNA code and released in a cellular environment to fold. *Folding is a self-assembly process, in which protein sequence spontaneously forms a unique native state*. Experiments of Anfinsen on the protein ribonuclease in '50s and '60s showed that a protein can be repeatedly unfolded and refolded back to the native state without any outside input. This implies that all "instructions", which a sequence needs to reach the native state, are encoded in a sequence itself. *It also follows that the native state is, in all likelihood, a global free energy minimum under given external conditions.* This idea constitutes the thermodynamic hypothesis in protein folding.

The problem of uncovering the mechanisms of protein folding is known as a folding problem. To understand folding problem let us consider the Levinthal's "paradox" first formulated in 1967 by Cyrus Levinthal. Consider a protein sequence of N=100 amino acids. Because atoms in amino acids can rotate with respect to each other, each amino acid can adopt, at a minimum, about 10 different conformations. (These conformations are due to the degrees of freedom associated with dihedral angles ϕ and ψ in Fig. 1.) The total number of conformations available for a chain is $C=10^N=10^{100}$. Assume that the rate of conformational sampling s is about 10^{14} conformations per second, i.e., the transition from one amino acid conformation to another takes about 10^{-14} sec. If protein folding is a random search process, in which all conformations must be "tried out", then the folding time scale is $\tau_F \sim C/s \approx 10^{80}$ years. This time scale is larger than the age of the Universe! Therefore, there must be some guided way to the native state or, in other words, some preferred folding pathways must exist.

"Old" view on protein folding: Levinthal's paradox may be resolved if one assumes that proteins fold through a series of well-defined intermediate structural states. The view was advanced through the study of a protein BPTI (bovine pancreatic trypsin inhibitor). This protein contains 56 amino acids, among which there are six cysteine (Cys) residues shown in Fig. 3. Cys residues have sulfur S atoms in their side chains, a pair of which can form disulfide bond S-S. Because these bonds are as stable as covalent bonds, the intermediate structures incorporating S-S bonds may be characterized in the experiments. Thomas Creighton in '70s showed that there is a well-defined sequence of states in the folding pathway of BPTI characterized in terms of S-S bond formation. For example,

with 60 percent probability the first bond to form is between cysteines at the positions 30 and 51 (a [30-51] bond). Late in folding a distinctive intermediate with two S-S bonds ([14-38],[30-51]) is detected. As a result the folding pathway may be represented as a sequence of intermediate states $\mathbf{I_n}$

$$\mathbf{U \Leftrightarrow I_1 \Leftrightarrow I_2 \Leftrightarrow .... \Leftrightarrow N,}$$

where **U** and **N** are unfolded and native states. Each of $\mathbf{I_n}$ is characterized by a particular combination of S-S bonds (Recommended reading: Kim and Baldwin *Annual Review of Biochemistry* **59**, 631 (1990)).

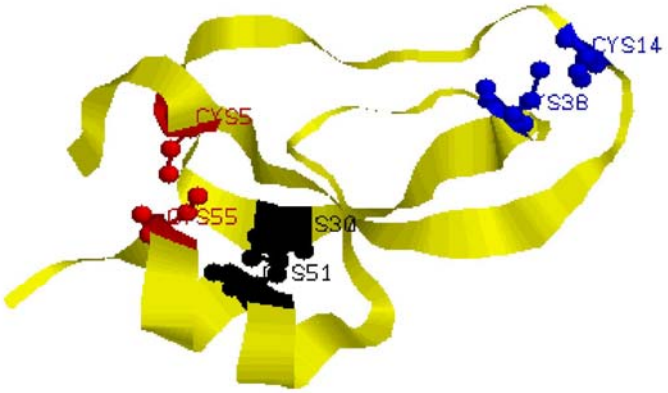


Fig. 3 Native structure of BPTI incorporates three disulfide bonds [5-55], [14-38], [30-51] between cysteine residues shown in red, blue, and black, respectively.

"New" view on protein folding: In equilibrium and kinetic experiments many small proteins (N<100) display very cooperative folding with no apparent intermediate states. Among these are CI2 studied by Alan Fersht in 1992 or ribonuclease A. The example of such cooperative equilibrium folding (more precisely, temperature induced folding) is shown in the following plot:

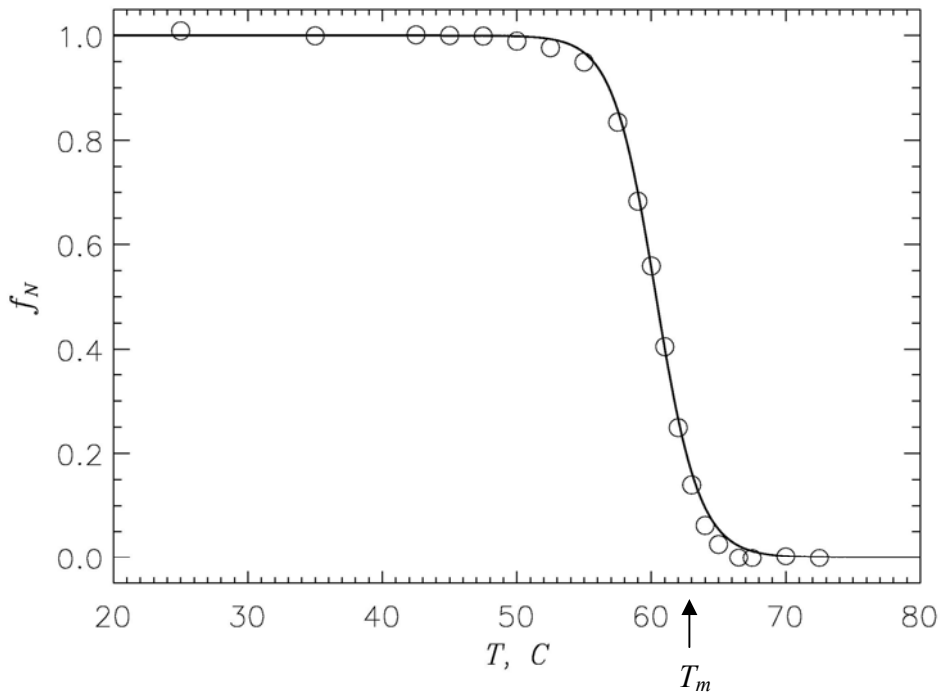


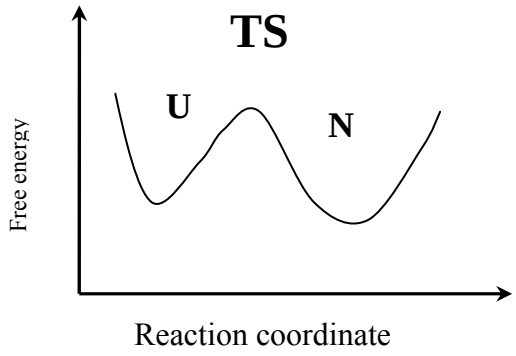
Fig. 4 Temperature induced folding of the protein RNase A is measured by the fraction of the native state f_N Sharp transition at the "melting" temperature T_m indicates a two-state folding, in which only unfolded and native states are populated. Circles correspond to experimental data.

Furthermore, folding kinetics of these proteins is usually single exponential. For example, the unfolded population of protein molecules decays exponentially with time as e^{-t/τ_f}. These observations taken together suggest that their folding largely involves only two states, native and unfolded, and can be represented as

$$\mathbf{U} \Leftrightarrow \mathbf{N}$$

The "new" view combines experimental observations of two-state folding with statistical mechanics concepts, such as polymeric nature of proteins, existence of the unique native state, multidimensionality of conformational space, and energy landscape perspective. The notion of energy landscape perspective lies in the center of the "new view".

Energy landscape of two-state folding proteins is visualized as a folding funnel (Fig. 5). The vertical axis shows the potential energy of a protein and the horizontal axis corresponds to folding reaction coordinate. The width of the funnel can be associated with the protein entropy. Myriads of unfolded states are located at the edges of the funnel, where potential energy and entropy are large. The native state is found at the funnel's bottom and has minimal energy and almost zero entropy. Folding, therefore, proceeds via a trade-off between energy and entropy – as protein moves down the energy landscape its energy decreases, but so does its entropy. Usually the gain in energy during initial stages of folding is not large enough to compensate for the loss in entropy that leads to existence of free energy barrier as shown.



In this figure transition state corresponds to the free energy folding barrier between native **N** and unfolded **U** states. Most importantly, folding takes place by different routes connecting unfolded and native states. According to Chan and Dill "folding is … a parallel flow process of an ensemble of chain molecules. [It] is seen … more like the trickle of water down mountainsides of complex shapes".

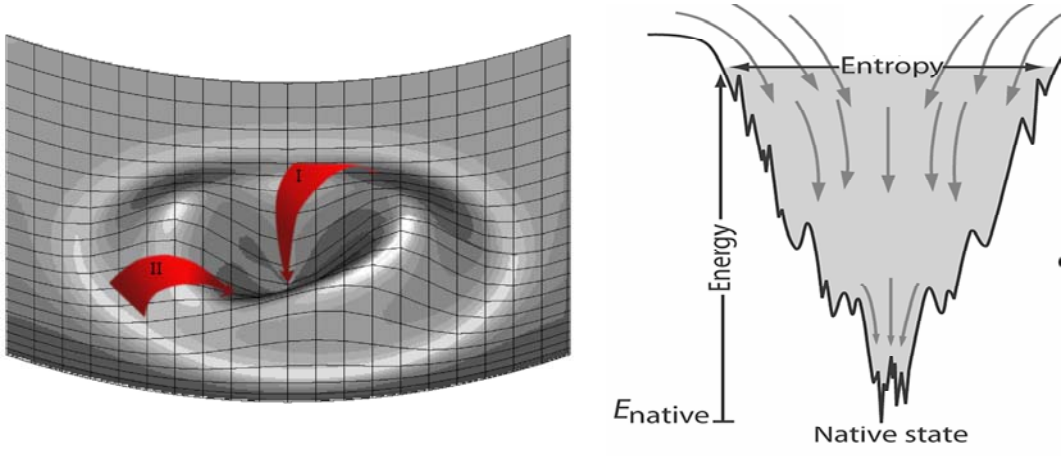


Fig. 5 Conceptual drawing of the free energy funnel (left panel) and its energetic version (right panel). The (free) energy is minimal at the bottom of the funnel, which corresponds to the native state. Energy barriers are shown as ripples in the energy landscape. Unfolded states are localized near the wide gulley in front of the folding barrier. Narrowing of the funnel closer to the native state represents the decrease in entropy, which is compensated by the decrease in energy (i.e., gain in attractive interactions). Red arrows indicate various routes to the native state.

"Ensemble" consequence of energy landscape perspective:

1. Many diverse parallel microscopic routes to the native state exist (tracks I and II in Fig. 5). Folding pathway is an ensemble of multiple microscopic folding trajectories. Microscopic folding routes may be grouped into several macroscopic pathways.

2. Any state except native should be viewed as an ensemble of microstates (i.e., microscopic conformations). This applies to unfolded or intermediate (if any) states.

***Homework**: read the article by Dill and Chan "From Levinthal to pathways to funnels"* (*Nature Structural Biology **4**, 10 (1997)).*

Illustration using lattice model: Lattice model (LM) is a very crude caricature of a real protein, but it retains several most important protein features, such as chain connectivity, steric interactions, and heterogeneity of amino acids and interactions between them. Basic principles governing protein folding may be gleaned from LM. Consider a two-dimensional LM, in which amino acids are confined to the lattice sites on a plane (see Fig. 6) Assume that there are only two types of amino acids – hydrophobic (H, filled circles) and polar (P, open circles), and the total number of amino acids is N=13. Assign attractive energy only to HH interactions, say, -1 (in the units of kT). The energy of a conformation E_p is then determined by the total number of nearest neighbor HH contacts, whereas mixed or PP contacts are assumed to have zero energy. The contacts, which are present in the native state **N**, are called native, and all others are considered as non-native. There are five contacts in the native state **N** indicated by dashed lines, so E_p=-5 (the bottom structures in Fig. 6). Let us follow some of the trajectories obtained in Monte Carlo simulations.

Folding starts with the unfolded state **U**, which has no HH contacts and, therefore, its energy E_p=0. Fig. 6 shows just one of such states. Soon the intermediate $\mathbf{I_1}$ is formed with the E_p= -2, when two local HH native contacts [2-5],[5-8] are established. Their formation does, however, block further progress to the native state, because it is not possible to reach **N** from $\mathbf{I_1}$ using the set of Monte Carlo moves. Therefore, partial unfolding of these contacts is needed and folding is essentially "restarted" from the other sequence end, when $\mathbf{I_2}$ is formed (local HH native contacts [8,13],[10,13] followed by the formation of long-range contact [2,13]). Transition from $\mathbf{I_1}$ to $\mathbf{I_2}$ involves crossing the energy barrier of, at least, 1. Intermediate $\mathbf{I_2}$ can be easily converted into **N** by forming the last missing native contacts [2,5] and [5,8].

Analysis of folding trajectories showed that occasionally non-native HH contacts form. For example, after $\mathbf{I_1}$ the intermediate $\mathbf{I_2^{NN}}$ occurs instead of $\mathbf{I_2}$. $\mathbf{I_2^{NN}}$ contains one non-native HH contact [5,10], which must be disrupted in order to reach **N**. Consequently, the transition from $\mathbf{I_2^{NN}}$ involves crossing the energy barrier of at least 1.

Lessons from LM studies:

1. Folding starts with local interactions;

2. Folding is not sequential process, in which the number of native contacts monotonically grows till the maximum number as in the native state. Occasionally folding process experiences "setbacks", when number of native contacts decreases.

3. Folding energy landscape is rough as transitions between the states may involve energy barriers (such as $\mathbf{I_1 \rightarrow I_2}$). Partial unfolding is required to overcome such topological barriers. It is important to consider connectivity between the states, i.e., accessibility of one state from another. For example, transition $\mathbf{I_1 \rightarrow I_2}$ is not possible without partial unfolding of $\mathbf{I_1}$.

4. Energy barriers may originate from topological constraints, i.e., due to the difficulty of rearranging a protein conformation without breaking covalent links between amino acids. In addition, non-native interactions contribute to roughness of energy landscape (intermediate $\mathbf{I_2^{NN}}$).

5. Folding trajectories can partition into fast and slow. This is a natural consequence of multiple folding routes. This may lead to multiphase folding kinetics.

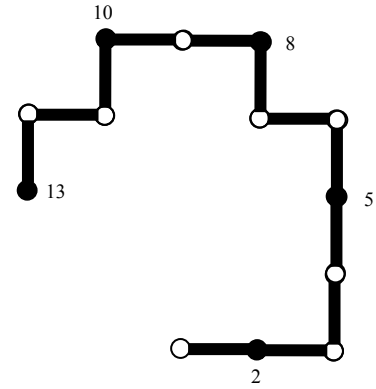
III. Folding mechanisms

Protein folding is mostly driven by hydrophobic interactions, formation of salt bridges and hydrogen bonds. Compaction of the chain is primarily due to non-specific hydrophobic collapse, which tends to bury hydrophobic residues in the interior of protein structure. Hydrogen bonds determine, to a large extent, secondary (local) structures in proteins. Salt bridges and steric interactions are more specific and define often the details of tertiary folds. For example, packing of side chains in a tight environment of native states is generally based on steric effects. There are several largely complimentary theories describing the mechanism of folding.

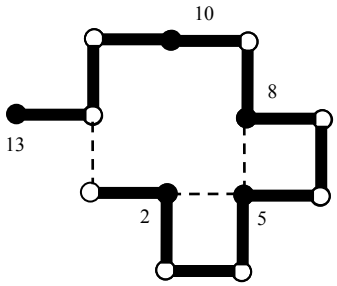
Hierarchic mechanism postulates that local interactions form first. Secondary structure elements, such as helices and strands, although having marginal stability, may be formed on their own without the help of long-range interactions. Native structure is assembled from the "preformed" pieces of secondary structure. There is considerable body of evidence that large proteins, such as α-lactalbumin, barnase, apomyoglobin, tendamistat fold (at least, at early stages) via hierarchic mechanism (Recommended reading: Rose and Baldwin *Trends in Biochemical Sciences* **24**, 26 (1999)).

Nucleation-condensation mechanism does not assume that folding is sequential. It suggests that local and tertiary structures are repeatedly formed and disrupted in search of a folding nucleus. As soon as a nucleus forms, i.e., the "right" combination of secondary and tertiary interactions is established, folding of the entire native structure takes place very fast. Therefore, nucleation-condensation mechanism postulates that there is a set of interactions in a protein, formation of which provides necessary and sufficient condition for fast folding. Nucleation-condensation mechanism is consistent with two-state folding of small proteins (such as CI2, SH3 domains etc) (Recommended reading: A. Fersht *Current Opinions in Structural Biology* **7**, 3 (1999)).

Fig. 6 Folding Pathways

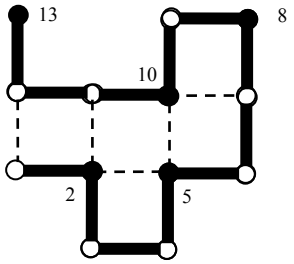


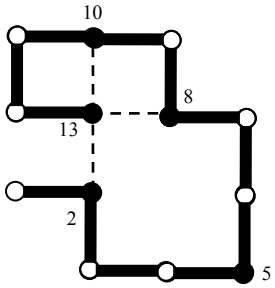
U: E_p=0



I_1: E_p = -2

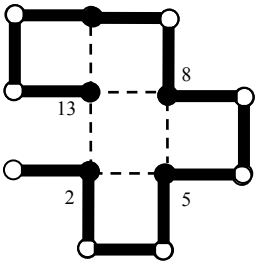
I_2^{NN}: E_p = -2

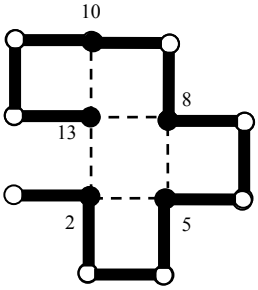




I_2: E_p = -3

N: E_p = -5





N: E_p = -5

Wild-type proteins are naturally designed and optimized for fast and reliable folding. Usually no external help, for example, from molecular chaperones, is need for successful monomeric folding. Note that random protein sequences, in which amino acids are placed at random, do not fold and often even do not have a unique native structure. Therefore, wild-type proteins are unique in their ability to fold fast to the native state and such sequences constitute a tiny fraction of all possible protein sequences. Protein folding is a very "delicate" process, which takes place in a very narrow window of external conditions (temperature, pH, denaturants, or presence of other molecules). For example, folding can be easily disrupted by interactions with other proteins (aggregation) that may prevent proteins to function properly and lead to various often fatal diseases.

IV. Folding timescales

The time scales for folding vary in a wide range. Formation of contacts between amino acids, which are close along the sequence (e.g., 3-4 amino acids apart), takes place within few nanoseconds. However, formation of interactions between more distant amino acids (e.g., 6 to 10 amino acids apart) takes up to a microsecond. Generally, α-helices are formed on a time scale of 10^2 ns, while β-hairpins form within approximately 10 μs. Folding time scale for proteins ranges from ~4 μs (35-residue villin headpiece subdomain) to hundreds of milliseconds (typically, large proteins, such barnase or lysozyme) (Recommended reading: S. Jackson *Folding & Design* **3**, R81 (1998)). It is also important to recognize that two time scales, for collapse and folding, characterize assembly of the native state. Collapse time scale τ_c is associated with general compaction of a protein and folding time scale τ_F is related to the formation of native interactions. For small two-state proteins τ_c and τ_F are almost equal, but for large proteins $\tau_c << \tau_F$. For example, for cytochrome c $\tau_c \approx 100$ μs, while $\tau_F \approx 400$ μs.

V. Computer simulations of folding

Computer simulations have played an important role in the study of protein folding. Simple lattice models revealed general physical principles governing folding, such as multiple folding pathways, funneled shape of energy landscape, etc. All-atom simulations were instrumental in studying the dynamics on short time scales, up to nanoseconds. These studies examined the structural changes of protein active sites, effects of denaturants and pH. All-atom simulations are routinely used to unfold proteins at highly denatured conditions. Recently, all-atom simulations start to approach the time scale of folding of peptides and small proteins, such as 23-residue BBA5 protein (*Nature* **420**, 102 (2002)).