# Chapter 2
# Classes of Protein and Their Folding

## 2.1 Introduction

The protein is functionally active in its three-dimensional native state. Almost all cells' machinery is based on the involvement of number of proteins. Structural code of proteins is hidden into the primary structure containing sequence of amino acids. It has always been a topic of research to investigate the mechanism of determining proteins' three-dimensional details from its primary structure. Studies on protein folding have shed various important facts; *viz*. the internal core of globular protein is formed by hydrophobic amino acid residues which are held together by van der Waals forces while surface is dominated by charged and polar side chains. Native state of the protein is the most active and stable state with a specific conformation determined by polypeptide backbone and intermolecular interactions. Native state of the protein has lowest free energy due to hydrophobic and electrostatic interactions as well as hydrogen bond energy [65]. Further lowest conformational entropy due to constrains imposed by $\psi$ and $\varphi$ bonds of main and side chains [66].

Protein folding is a complex issue in case of multidomain proteins (containing more than one domain). Generally proteins with amino acids more than 200 belong to the category of multidomain. They are found to have independent folding of each domain behaving like independent whole protein. However, presence of other domains is helpful in imparting stability to the respective domain. It has been found that rate of folding is lowered when folding pathway of single isolated domain is studied. Rate limiting step in folding of multidomain protein is the interfacial interactions between individual domains [67, 68]. Folding kinetics follows an initiation fast phase followed by intermediate state lacking various properties of native protein, more labile to proteolysis and lacks catalytic activity and finally last stage with very slow rate of folding. It is during last stage when pairing of already folded domains present in the protein takes place. Rate of folding is found to be inversely related to the solvent viscosity [69]. High protein concentration as found in cellular environment has been found to be helpful in domain pairing but it usually promotes dimer formations rather monomer [70]. It has been found that ligand binding is also



A. Dwevedi, *Protein Folding,* SpringerBriefs in Biochemistry and Molecular Biology,
DOI 10.1007/978-3-319-12592-3_2

one of the factor responsible for proper protein folding as found in various cases (troponin C site III, Arc repressor, Trp repressor, p53, HIV gp41) [71].

Molecular chaperones are the major proteins guiding protein folding and proper assembly of macromolecular structures in prokaryotic as well as eukaryotic cells. The most important class of chaperons is heat shock proteins, *viz*. Hsp10s, Hsp40s, Hsp60s, Hsp70s and Hsp90s. Efficient protein folding is the pre-requisite of proper functioning of cell machinery which depends on coordinated functioning of chaperones, chaperonins, and various auxiliary cofactors [72]. In some cases, this machinery fails due to which synthesized proteins could not be folded properly leading to their aggregation. The failure of a specific protein to adopt its native and active state is called "protein misfolding". Protein misfolding has wide range of pathological implications due to loss of normal cellular functions. Misfolded proteins form fibrillar aggregates called amyloid fibrils which are resistant to ubiquitin–proteasome degradation system due to which their accumulation occurs inside/outside the cell [73]. Further, aggregation and amyloid formation of a protein may also be promoted due to various point mutations which destabilize its three-dimensional structure and induce amyloidogenic conformations. Formation of the correct native-like disulfide bridges (between a pair of cysteine residues) is another factor in protein misfolding. Greater the number of disulfide linkages, higher is the chance of misfolding due to errors in disulfide pairing [74].

## 2.2 Structural Classes of Protein (Fig. 2.1)

(www.swissmodel.expasy.org, www.wikipedia.org/wiki/structure, www.cryst.bbk.ac.uk)

a. All-$\alpha$

- *Lone helix*: Small proteins containing a single helix, *viz*. alamethicin (transmembrane voltage gated ion channel).
- *Helix-turn-helix motif*: Two helices lying antiparallel connected by short loop, *viz*. RNA binding protein Rop.
- *Four-helix bundle*: Bundle of four helices connected by three loops. Here, interfaces in between the helices are hydrophobic while surface is hydrophilic. They are present in photosynthetic reaction centre, membrane spanning region of G-protein coupled receptors, steroid-binding proteins like uteroglobin, ferritin, cytokines (interleukin-2, granulocyte-macrophage colony-stimulating factor, GM-CSF), DNA binding proteins (usually transcription factors), etc. They are also present in globin fold containing cluster of two bundles, each of four helices. Globin fold is also called "Greek key helix bundle" due to its topological similarity.
- *Helix-helix packing*: $\alpha$-helices are packed in such a way that they have complementary interfacial regions with buried side chains. Their surface is not smooth, characterized by grooves and ridges with each ridge are at 26° from the main axis, for example: carboxypeptidase A, flavodoxin, subtilisin, etc.

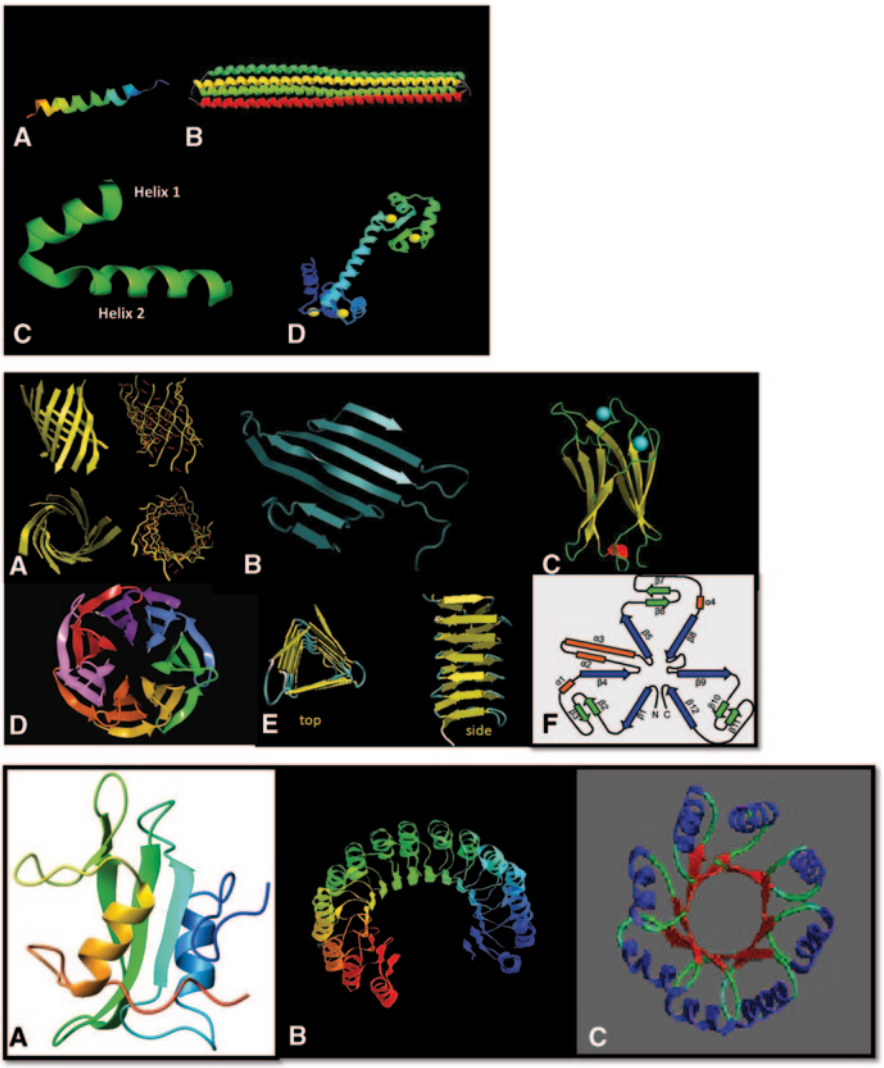


**Fig. 2.1** Structural representation of protein models belonging to all-$\alpha$, all-$\beta$ and mixed class. **(I) all-*α***, **A**: lone helix, **B:** Four-helix bundle, **C:** Helix-turn-helix motif, **D:** Helix-helix packing. **(II) all-*β***, **A:** $\beta$-Barrels, **B:** Up and down antiparallel $\beta$-sheets, **C:** $\beta$-Sandwiches, **D:** $\beta$-Propellers, **E:** $\beta$-Helix, **F:** $\beta$-Trefoils. **(III) Mixed class**, **A:** $\alpha$+$\beta$, **B:** $\alpha/\beta$ horseshoe, C: $\alpha/\beta$ barrels. (Adapted from www.swissmodel.expasy.org, www.cryst.bbk.ac.uk)

b. All-$\beta$

- *$\beta$-Sandwiches*: They are also called immunoglobulin fold containing $\beta$-strands forming two sheets which are packed like sandwich. The packing of two sheets is either aligned (mean angle between the two sheets is ~30°) or orthogonal (two sheets are at 90°). The two sheets are independent which are linked by residues not in $\beta$-sheet conformation. Here side-chains are not fixed

at any angles to the interface. Examples are: superoxide reductase, clathrin adaptor, transglutaminase, $\alpha$-amylase inhibitor, etc.

- *$\beta$-Barrels*: Domain present in the protein containing antiparallel $\beta$-sheet without any fixed arrangement of $\beta$-strands, for example streptavadin and porin.
- *Up and down antiparallel $\beta$-sheets*: Here antiparallel $\beta$-strands making sheet are connected by loops of adjacent strand resembling to Greek key. Three up-and-down $\beta$-strands are connected by hairpins, followed by fourth strand lying adjacent to the first. Examples are: plastocyanin and $\gamma$-crystallin.
- *$\beta$-Propellers*: This fold is a superbarrel containing six, four-stranded stranded antiparallel sheets with up-down topology, for example calcium dependent *Bacillus* phytase.
- *$\beta$-Trefoils*: It has an ~3-fold axis of symmetry, for example cytokinin interleukin-1$\beta$.
- *$\beta$-Helix*: This fold has resemblance to helical topology with $\beta$-strands wound round the structure, for example, monomeric left handed $\beta$-helix antifreeze protein from spruce budworm.

c. Mixed Class ($\alpha/\beta$, $\alpha+\beta$)

- *$\alpha/\beta$:* This fold is most commonly found in number of proteins which contains repeating $\beta$-$\alpha$-$\beta$ supersecondary units (right handed) with outer layer composed of $\alpha$-helices and central core of parallel $\beta$-sheets. The $\alpha$-helices and $\beta$-strands are parallel to each other, while $\alpha$-helices are antiparallel to the strands. This fold is also called Rossman Fold named after Michael Rossman. Many enzymes of glycolysis, various cytosolic proteins, and nucleotide binding proteins have this characteristic fold.

  1. *$\alpha/\beta$* horseshoe: As the name represents, they look like an open horseshoe containing a curve made by repeating units *$\alpha/\beta$* with parallel $\beta$-sheet while $\alpha$-helices are at the surface of the curve. The $\beta$-strands are parallel to the central axis while they are slightly slanted with respect to each other. For example, placental ribonuclease inhibitor.
  2. *$\alpha/\beta$* barrels: Here sequence of eight $\beta$-$\alpha$ with first strand hydrogen bonded to the last strand, forming a barrel-like structure. The fold is not open rather than closed like barrel with $\alpha$-helices situated on only one side of the $\beta$-sheet. The most important example: triose phosphate isomerase.

- *$\alpha+\beta$:* They contain significant $\alpha$ and $\beta$ secondary structural elements, not having any specific topology. Example: cysteine proteases (papain and actinidin), DNA-binding protein, microbial ribonucleases, lysozome, chalcone isomerase, ribonuclease-H, carbonic anhydrase, serine protease inhibitor, thymidylate synthase etc.

## 2.3 Correlation of Protein Folding with Structural Classes

Protein folding, particularly *in vitro* has been a most interesting aspect toward physicists, chemists and biologists. The ability of proteins to fold spontaneously immediately is the most crucial fundamental problem being solved since 1960s. Protein attains its native state *in vivo* with the help of various chaperones just after its synthesis on ribosome. *In vitro* protein folding has been given more priority in research with its association of various interesting facts, *viz*. how the chain can find its most stable structure within seconds, prediction of three-dimensional structure from amino-acid sequence of the protein etc. However, both *in vivo* and *in vitro* protein folding have emphasized that the native state is the most thermodynamically stable state with $\Delta G = 0$. Protein can take zillions of possible conformations, once the right stable conformation is achieved then deviation of ~1 Å can strongly increase the chain energy by several folds [75]. There must be some specific folding pathway which propels the unfolded protein in that particular direction making the process so fast. Molecular simulations studies using lattice models of protein chains have shown that protein folding is initiated by nucleus formation with rate of folding dependent on the size of the protein while whole process is under thermodynamic control. In earlier days, it was believed that the protein folding is proceeded by formation of nucleus by *N*-end followed by wrapping of remaining chain around nucleus which was proven incorrect in later days. Subsequently, various theories have been proposed to elucidate the mechanism of protein folding as discussed briefly here (www.wikipedia.org/protein_folding):

a. *Nucleation/growth model:* The rate-limiting step during protein folding is occurrence of nucleation (formation of smaller structural units). Once nucleation begins, it generates number of nuclei which fastens the protein folding by several folds. This model could not be fitted in various folding experiments observing folding intermediates.
b. *Diffusion–collision–adhesion model:* Protein folding is brought up by repeated diffusion and collisions of microdomains (enriched with hydrophobic clusters containing secondary structures) leading to generation of larger units. Here, rate of diffusion is the determining factor for rate of protein folding.
c. *Framework model:* It states that the protein folding is hierarchical which begins with formation of secondary structures followed by tertiary, accompanied with inter- and intra-chain interactions. Here, formations of secondary structures are the rate determining step during protein folding.
d. *Hydrophobic collapse model*: It is based on the fact that protein folding is brought up by rapid collapse of hydrophobic clusters followed by formation of secondary structures. It has not been experimentally verified whether formation of secondary structures is the initiating or intermediate step during protein folding.
e. *Jigsaw puzzle model:* This model states that different proteins have different route of folding pathway similar to the fact that there are multiple ways of solving jigsaw puzzle. This model is well suited for energy landscape view stating

that native structure of protein is at global minimum while unfolded state at global maximum taking a shape of funnel with each molecule following different microscopic route from top to bottom.

f. *Nucleation-condensation model:* This model agrees with both the framework and the hydrophobic collapse mechanisms. It states that long-range as well as native hydrophobic interactions are important in the formation of transition state which imparts stability to formed-secondary structures.

Folding funnel model has gained much popularity with respect to other models for description of fast protein folding processes. Folding funnel as hierarchical folding has found that rate of protein folding is in minutes rather in astronomical numbers as predicted from Levinthal calculations. However, none of the models could explain the mechanism of protein folding. This section will be dealing with rate of protein folding with respect to different structural classes under *in vitro* conditions. The important factors are: size, amino-acid composition, chain length, native-state topology. Rate of protein folding has been considered to be an important aspect as it will give an insight into underlying folding mechanisms.

Protein unfolding under *in vitro* conditions is done using various denaturants (urea, guanidium hydrochloride), higher temperature, extreme pH, solvents etc. Based on the thermodynamics and kinetics studies of protein unfolding/refolding, monomeric and dimeric models are reported. Monomeric models comprise of native ($N$) and denatured ($U$) states present at the beginning and completion of reaction. They are of two types: two states and three-states.

1. Simplest is the two-state:

$$N \underset{\longleftarrow}{\overset{K_{eq}}{\longrightarrow}} U$$

where, $K_{eq}$ represents the equilibrium constant of the reaction.

2. Three-state: Here native ($N$) protein unfolds through a partially structured intermediate ($I$) as shown:

$$N \underset{\longleftarrow}{\overset{K_1}{\longrightarrow}} I \underset{\longleftarrow}{\overset{K_2}{\longrightarrow}} U$$

$K_1$ and $K_2$ are the equilibrium constants of the two reactions.

Dimeric models comprise of more than one state during beginning and completion of the unfolding/refolding reactions. Various types of dimeric models are:

1. Two-state: Native dimer ($N_2$) end with two unfolded monomers ($2U$)

$$N_2 \underset{\longleftarrow}{\overset{K_{eq}}{\longrightarrow}} 2U$$

2. Three-state: There are two ways by which protein unfold/refold using three-states,
   - First case: Monomeric intermediate ($2I$) is populated between $N_2$ and $2U$

$$N_2 \underset{}{\overset{K_1}{\longleftrightarrow}} 2I \underset{}{\overset{K_2}{\longleftrightarrow}} 2U$$

– Second case: Dimeric intermediate ($I_2$) is populated between $N_2$ and $2U$

$$N_2 \underset{}{\overset{K_1}{\longleftrightarrow}} I_2 \underset{}{\overset{K_2}{\longleftrightarrow}} 2U$$

$K_1$ and $K_2$ are the equilibrium constants for each transition.

3. Multiple state: It exists when there are different forms of intermediates between states $N_2$ and $2U$. Following are the equations describing various cases.

$$N_2 \underset{}{\overset{K_1}{\longleftrightarrow}} 2I^x \underset{}{\overset{K_2}{\longleftrightarrow}} 2I^y \underset{}{\overset{K_3}{\longleftrightarrow}} 2U$$

$$N_2 \underset{}{\overset{K_1}{\longleftrightarrow}} I_2 \underset{}{\overset{K_2}{\longleftrightarrow}} 2I \underset{}{\overset{K_3}{\longleftrightarrow}} 2U$$

$$N_2 \underset{}{\overset{K_1}{\longleftrightarrow}} I_2^x \underset{}{\overset{K_2}{\longleftrightarrow}} I_2^y \underset{}{\overset{K_3}{\longleftrightarrow}} 2U$$

Dimeric intermediates ($I_2$, $I_2^x$, $I_2^y$), monomeric intermediates ($2I$, $2I^x$, $2I^y$), unfolded monomers (2U), and $K_1$, $K_2$ and $K_3$ are the equilibrium constants for the three transitions.

It has been found that rate of unfolding/refolding depends on various factors: native-state topology [*viz*. all-$\alpha$, all-$\beta$, mixed class ($\alpha+\beta$ or $\alpha/\beta$)], size, and amino acid composition [76, 77]. As the studies on protein unfolding/refolding are progressing, various facts have been elucidated regarding correlation of rate of unfolding/refolding with various factors including structural class of proteins. However, these studies are mostly limited for two-state monomeric models with very few reports on three-state monomeric models. There are no reports on any dimeric models till date due to complexity of the reactions. Following discussion will be on correlation of folding rate with respect to two-state monomeric models only.

Brief description of the terms are given below which will be used in following text to understand the correlation of protein folding rate with various factors.

- Contact order (CO): measure of inter-amino acid contacts in the native state of protein structure. It is estimated as the average sequence distance between residues forming native contacts within the folded protein divided by the total length of the protein. Higher the value of contact orders greater would be the time of protein folding.
- Relative contact order (RCO): measure of the relative interactions of local *versus* non-local noncovalent interactions.
- Absolute contact order (ACO): average sequence separation of contacting residues.

- Long range order (LRO): measure of amino acid interactions with the distance away from more than four amino acids.

Structure topology is the main determinant of the folding rate of small proteins following two-state kinetics. They are independent of chain length as found from experimental and theoretical reports [78–80]. In case of larger proteins with observable intermediates (monomeric models), chain length is the main determinant of their folding rates. It has been found that there is no correlation of three-state kinetics following proteins with RCO while it is logarithmically related in two-state kinetics following proteins [81]. The analysis of 56 non-redundant two-state proteins with chain length varying from 16 residues (the *C*-terminal $\alpha$-hairpin peptide of the B1 domain of protein G) to 322 residues (4 $\alpha$-helix bundle of the VlsE antigen protein (PDB ID: 1L8W) was done to determine the correlation of folding rate with factors (chain length, amino-acid composition and surface topology) [78].

Before going into the detailed analyses of the report, following are given relation of RCO and LCO with factors [82]:

$$RCO = \frac{1}{N_c L} \sum_{\substack{contacting \\ atomsi,j}} d_{ij}$$

where,

$L$ chain length
$N_c$ total number of contacting atoms (using a 6 Å distance threshold)
$d_{ij}$ number of residues separating those two residues to which atoms $i$ and $j$ belong.

$$LRO = \frac{1}{L} \sum_{\substack{contacting \\ residuesi,j}} n_{ij},$$

where, $n$ is the number of residues separating those two residues to which atoms $i$ and $j$ belong with their separation always less than 12 [83].

Absolute contact order (ACO) is correlated to RCO as:

$$\mathrm{ACO} = \mathrm{L} \times \mathrm{RCO},$$

Further, two residues are considered to be in contact if the closest distance between their $C_\alpha$ atoms is $\leq 8$ Å.

Figures 2.2a, b, c and d have shown the correlation of between logarithmic folding rates, $\ln k_f$ and basic structural and topological parameters for proteins belonging to three structural classes (all-$\alpha$, all-$\beta$ and mixed class). Figure 2.2a, shows correlation of total chain length ($L$) and rate of folding ($\ln k_f$). Straight line was observed in case of all-$\alpha$ and all-$\beta$ with correlation coefficients as –0.80 at $P$-value of $5.7 \times 10^{-4}$ and –0.80 at $P$-value of $6.0 \times 10^{-5}$, respectively. In case of mixed-class

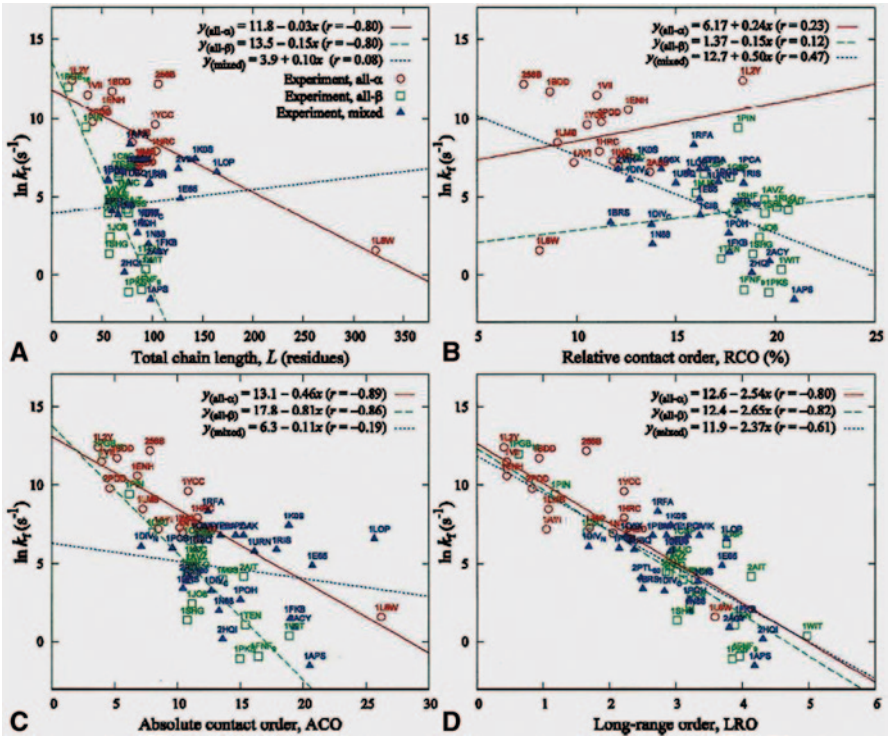


**Fig. 2.2** Correlation between natural logarithmic folding rates ($\ln k_f$) of different proteins (mentioned with their PDBID) with their basic structural and topological parameters. **(A):** $\ln k_f$ *versus* protein chain length, L, **(B):** $\ln k_f$ *versus* relative contact order, RCO, **(C):** $\ln k_f$ *versus* absolute contact order, ACO, **(D):** $\ln k_f$ *versus* long-range order, LRO. (Adapted from [78])

proteins, there is almost no correlation between chain length and rate of folding ($\ln k_f$), showing that their folding mechanism is different and more complex than pure class. These results are concluded with the fact that the chain length is linearly correlated to folding rates in case of all-$\alpha$ and all-$\beta$ two-state proteins while it's independent in mixed class. Figures 2.2b, c and d, show the correlation of RCO, ACO, and LRO with rate of folding in pure class (all-$\alpha$ and all-$\beta$) and mixed class. The RCO has significant impact on $\ln k_f$ for small two-state proteins of all three structural classes in the range when chain length was ~100–250 residues. However, it does not fit well when protein is too small, *viz*. all-$\alpha$ class, 20-residues [Trp-cage miniprotein construct, TC5b (PDBID: 1L2Y) of the 322-residue VlsE protein (PDB: 1L8W)]; all-$\beta$ proteins, 16-residues [*C*-terminal $\beta$-hairpin of protein G of the 34-residue subdomain of peptidyl-prolyl *cis-trans* isomerase (PDBID: 1PIN)] with completely disrupted correlations of RCO with their folding rate. On the other hand, ACO shows strong correlation of high statistical significance with the folding rates of pure class indicating that both chain length and structural topology have significant affect on folding rate of two-state proteins. LRO has been found to be uniform

topological descriptor of folding rates of all the three structural classes. There is strong correlation with correlation coefficients as: $r_{\text{all}-\alpha}=-0.80$, $r_{\text{all}-\beta}=-0.82$, and $r_{\text{mixed}}=-0.61$. It can be interpreted that the rate-limiting step is the formation of $\beta$-sheet and loop structures (i.e., formation of contacts that are long range in sequence, whose rate is limited by cooperative diffusion) due to which rate of folding is slower in case of all-$\beta$ and mixed class with respect all-$\alpha$ class. Higher the rate of formation of secondary structures, greater would be rate of protein folding in all cases supporting hierarchical view of folding [78, 84, 85].

*Correlation of free energy with rate of folding*: The free energy ($\Delta G$) of the native globular protein which is not covered with disordered loops containing L residues can be related as [86]:

$$\sigma = 2.3RT \text{ x } 0.33 \approx 0.7RT$$

where,

| | |
|---|---|
| $g$ | free energy of one residue inside the globule |
| $\sigma$ | free energy lost by one residue on the globule's surface |
| $B_L L^{2/3}$ | number of residues at the surface of the native globule |

For the compact and spherical globular protein, $\sigma \approx 1/6 \times 1.2L^{2/3} \times 5RT + 4/6 - 1.2L^{2/3} \times 0.7RT \approx 1.5L^{2/3}RT$. which is only 7.6–8.7 % greater with respect to two-fold oblong or oblate ellipsoid, and $\sigma = 2.3RT \text{x} 0.33 \approx 0.7RT$, where $2.3RT$ is the average residue's energy lost upon protein denaturation at temperature $T$ [87].

Protein folded to its native states leads to $\Delta G = 0$, thus $g = -sB_L L^{-1/3}$. This suggests that the surface stability is an important parameter for achieving stable native state due to its direct interactions with the surrounding solvent responsible for number of conformational variations in the main as well as side chain. Further, if protein folding goes *via* spherical (the least unstable) intermediate structures, free energy of the fastest pathway is given as [88]:

$$\Delta G_{\#} = \frac{4\sigma B_{sph} L^{2/3} (B_{sph} / B_L)^2}{27}$$

whereas, the folding nucleus size (central region of the protein) given as:

$$L_{\#} = \frac{8L(B_{sph} / B_L)^3}{27}$$

For a spherical central globule, $B_{\text{L}} = B_{\text{sph}}$ and $L_{\#}/L \approx 0.30$, free energy is given as:

$$\Delta G_{\#} / RT \approx \frac{L^{2/3}}{2}$$

http://www.springer.com/978-3-319-12591-6

Protein Folding
Examining the Challenges from Synthesis to Folded Form
Dwevedi, A.
2015, VII, 55 p. 6 illus., 4 illus. in color., Softcover
ISBN: 978-3-319-12591-6