Secondary and tertiary protein structure

I. Hierarchy of protein structure

Four levels in protein structural organization are commonly identified. *Primary structure* is a sequence of amino acids. *Secondary structure* is represented by regular local conformations of polypeptide chain, such as α-helix or β-strand. The combinations of two secondary structure elements are also sometimes referred to as secondary (or *supersecondary*) structure. The example is a β-hairpin formed by two adjacent β–strands. The entire 3D distribution of protein atoms is termed as *tertiary structure*. *Quaternary structure* describes the 3D arrangement of individual domains in large multidomain proteins.

The local (secondary) structure in proteins is conveniently characterized by Ramachandran plots, which display the distribution of allowed (ϕ,ψ) angles. Because of steric hindrance, relatively few areas of Ramachandran plot are actually populated. Fig. 1 shows the computation of (ϕ,ψ) angles distribution for 403 PDB X-ray crystallographic structures resolved with the accuracy of 2.0 Å or better (*Structure* 4, 1395 (1996)). The magenta contour line encloses the area, in which 98% of all non-glycine residues are found. In total, the enclosed area comprises merely 20% of the entire plot.

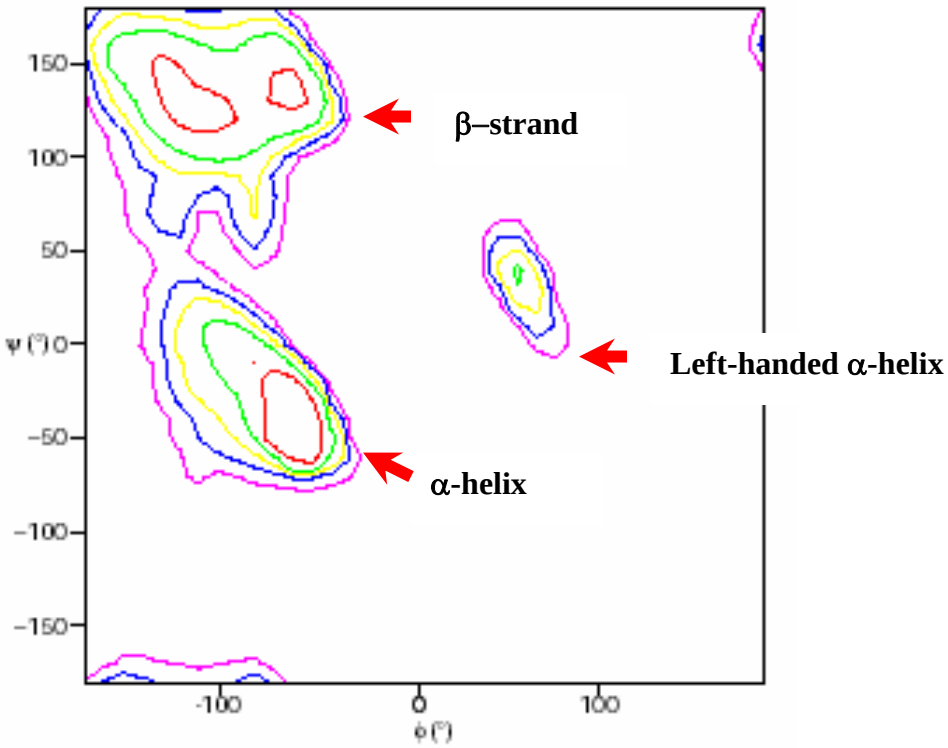

Fig. 1 Ramachandran plot of the accessible (ϕ,ψ) angles for high-resolution PDB protein native structures.

II. α-helix

The area of Ramachandran plot with ϕ~-60° and ψ~–50° corresponds to classical right-handed α-helix. The energetic stability of α-helix is related to a regular hydrogen bond (HB) pattern, in which CO group of the residue *i* makes a HB with the amide group NH of the *i*+4 residue as shown in Fig. 2. The ideal α-helix has 3.6 residues per turn (helical pitch) and typically spans from 10 to 15 residues in a protein sequence. Many proteins contain extensive α-helical structure, such as ACBP (Fig. 5) or myoglobin (Fig. 6)

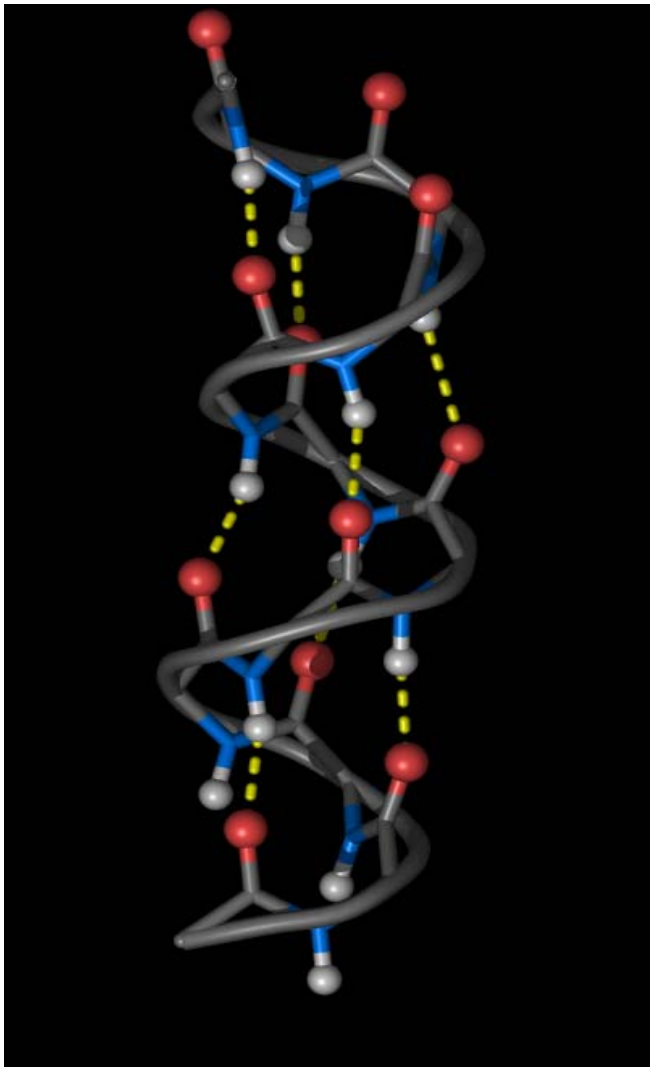

Fig. 2 The backbone trace of the α-helix. Hydrogen bonds between carbonyl oxygens O_i (in red) and amide groups NH_{i+4} (in blue/grey) are shown by yellow dashed lines. The backbone trace is given by a smooth grey tube.

The α-helical regions in protein conformations can be identified by calculating the distribution of (ϕ,ψ) dihedral angles. There is no universal definition of α-helical (ϕ,ψ)

angles. Rose and coworkers (*Proteins Structure Function Genetics* **22**, 81 (1995)) proposed to assign α-helix to a protein conformation, if -80° < ϕ < -48° and -59° < ψ < -27° (a "strict" definition). More inclusive ("broad") definition is suggested by Serrano and coworkers (*Proteins Structure Function Genetics* **20**, 301 (1994)). According to this definition the ϕ and ψ axes in Ramachandran plot are divided into 20 equal intervals to create a uniform grid covering the entire plot. The α-helix structure corresponds to the area enclosed by the polygon (-90°,0°),(-90°,-54°), (-72°,-54°), (-72°,-72°), (-36°,-72°), (-36°,-18°), (-54°,-18°), (-54°,0°).

The α-helix structure can also be defined based on characteristic (*i*,*i*+4) HB pattern as it is done in DSSP database. Table 1 demonstrates that several residues have a high propensity to form α-helix, such as Ala, Met, Glu, or Lys. Pro with the constrained side chain is especially poor helix former. In addition to HBs, α–helix may draw its stability from salt bridges and hydrophobic contacts.

Besides α-helix two other, special types of helices exist. Tight 3_{10}-helix has the characteristic (*i*,*i*+3) pattern of HBs, i.e., CO_i group forms a HB with the NH_{i+3} group. This helix has only three residues per turn and each HB spans 10 heavy atoms in a sequence. Tight packing of side chains in 3_{10}-helix is energetically unfavorable, therefore, it rarely extends by more than few residues or is stable in a solution as an isolated fragment. The typical values of ϕ and ψ angles for 3_{10}-helix are -50° and -25°, respectively.

A loosely packed π-helix has the (*i*,*i*+5) HB pattern and is wide enough to allow water penetration along its axis. Similar to 3_{10} helix π-helix is usually unstable without support of tertiary interactions. The typical values of ϕ and ψ angles are -60° and -70°, respectively. Both special helices are populating the fringes of α-helix region in the Ramachandran plot. Left-handed α–helix is unstable because of the L-chirality of amino acids.

III. β-sheet structure

In addition to α-helix β-strand local structure is another common secondary structure type in native proteins. β–strand corresponds to extended effectively planar conformation of atoms in a protein backbone. According to "strict" definition the dihedral angles in a β–strand are in the range -150°<ϕ<-90° and 90°<ψ<150°. The "broad" definition assumes that β–strand conformations of amino acid correspond to the region enclosed by the polygon (-180°,180°), (-180°,126°), (-162°,126°), (-162°,108°), (-144°,108°), (-144°,90°), (-50°,90°), (-50°,180°) (*Proceedings National Academy of Sciences* **96**, 9074 (1999)). A typical length of a β–strand is from 5 to 10 residues. Usually β–strands are not isolated, but participate in β–sheet structure based on extensive interstrand HB network. The examples of proteins containing β–sheets are immunoglobulin domains of titin (e.g., Ig27, Fig. 7), fibronectin domains (e.g., 10FnIII, PDB access code 1fnf), or transpheretin (PDB access code 2pab).

There are two possible geometrical arrangements of β–strands in β–sheets. In *antiparallel* β–sheets individual β–strands are oriented in antiparallel way and the typical (ϕ,ψ) angles are -140° and 135° (Fig. 3). An elementary antiparallel β–sheet is represented by a β–hairpin (Fig. 4), in which two β–strand are linked by a turn. If the total number of residues in the hairpin is *N*, then the HBs are formed between the *i* and *N*-*i*+1 residues. In antiparallel β–sheets HBs are in-registry, because they are formed between the residues, which are direct counterparts in the neighboring strands.

Parallel β–sheets are formed when β–strands are oriented parallel to each other (Fig. 3). For this type of β–sheet the typical dihedral angles are ϕ=-120° and ψ=115° and HBs are out-of-registry. The antiparallel β–sheets accommodate HBs slightly better than parallel ones. Several residues show a clear propensity to adopt β–strand structure, such as Val, Tyr, Phe, Ile, whereas others, such as Pro and Asp are rarely found in these conformations.

IV. Turns and loops

Turns correspond to short parts of sequence that make a sharp change in direction. Turns are typically found between two antiparallel β–strands (e.g., hairpin in Fig. 4). By definition, two residues, *i* and *i*+1, form a turn, if the distance between the residues *i*-1 and *i*+2 is less than 7Å. Usually, the turn region is flanked by HBs (Fig. 4). Loops (usually of up to five residues) are longer than turns and connect other types of secondary structure elements, such as helices and strands. Turns and loops are generally exposed to solvent. Pro, Gly, polar Ser and Asp are often found in these sequence regions (Table 1).

V. Tertiary structure

Tertiary structure of proteins is built from the secondary structure elements, such as α–helices, β–strands etc. Yet there is no clear boundary between tertiary and secondary structures. For example, hairpins or β–α–β motifs are considered simultaneously as the examples of supersecondary and tertiary structures. In general, protein structures are divided into four classes (for current statistics, see the Structural Classification of Proteins (SCOP) database website at scop.mrc-lmb.cam.ac.uk/):

1. α–helix class includes all the proteins, whose native states contain mostly α–helices usually wrapped around common hydrophobic core. Approximately, 22 % of proteins in PDB fall into this category

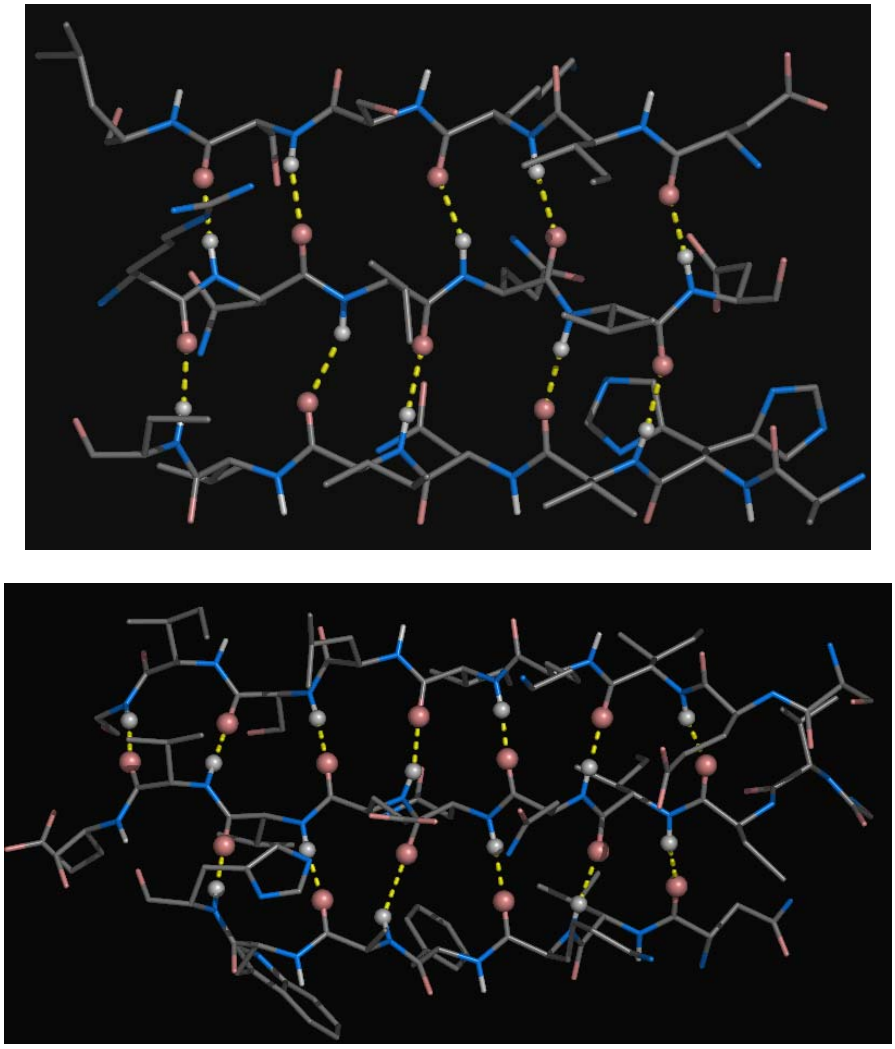

Fig. 3 Antiparallel (upper panel) and parallel (lower panel) β–sheets. Hydrogen bonds between carbonyl oxygens (in pink) and amide groups (in blue/grey) are marked by yellow dashed lines.

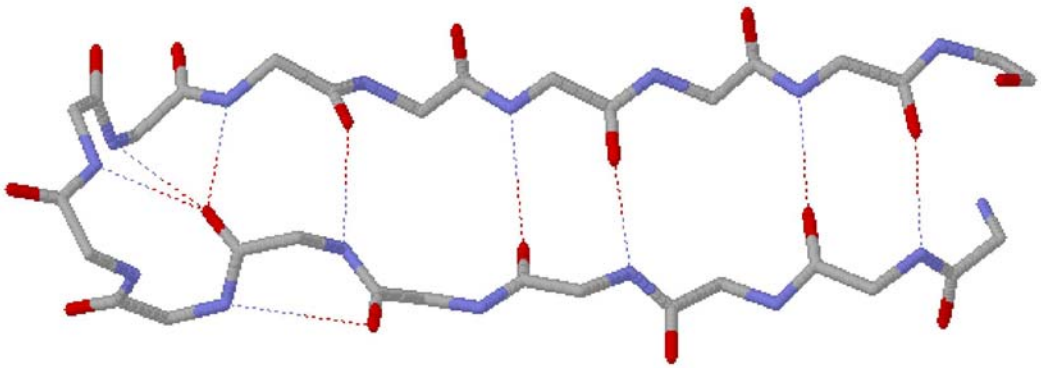

Fig. 4 Structure of β–hairpin from GB1 protein illustrates the elementary antiparallel β–sheet. Hydrogen bonds are shown by dashed lines between backbone oxygens (in red) and amide groups (in blue).

2. β–proteins contain mostly β–strands arranged into β–sheets, which are stabilized by HB network. These proteins typically have layered structure with the hydrophobic core. About 16 % of PDB proteins belong to this class.

3. α/β proteins contain alternating pattern of α– and β–structure (15 % of PDB database).

4. α+β proteins also contain mixed α– and β–structure, but it is spatially separated (29 % of PDB structures).

In terms of their further structural characterization classes are divided into folds. Proteins from a common fold share the same core secondary structure elements in the same arrangement (*Current Opinions in Structural Biology* **7**, 369 (1997)).

Proteins from α–helix class are divided into bundle, folded leafs, or hairpin array folds. The α–helix bundles are formed when several α–helices are wrapped around a common hydrophobic core. An example of such fold is a native structure of bovine ACBP (Fig. 5), which contains four α–helices.

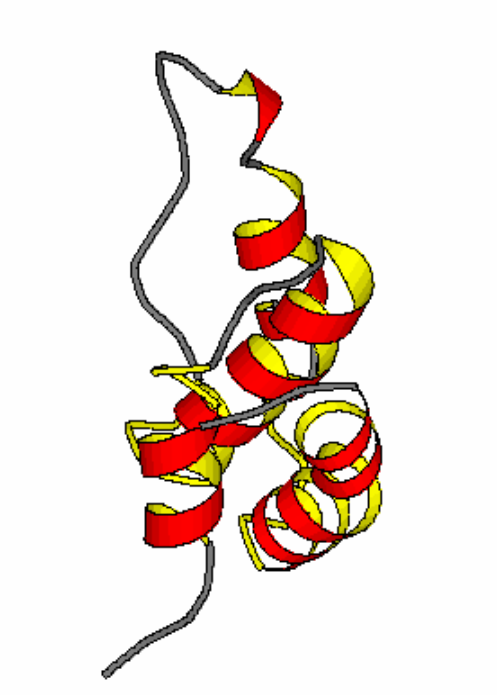

Fig. 5 Native structure of ACBP (access code 2abd) demonstrates the example helix bundle fold.

The folded leaf fold is present in the native state of myoglobin (Fig. 6), in which eight helices are arranged in multilayered structure. The tertiary structure of myoglobin is stabilized largely by hydrophobic interactions. These are just several examples of the α–class folds (see SCOP website for more complete information).

β–proteins are also subdivided into several folds. One of the most typical is a β–sandwich fold. This fold adopted by the native structure of immunoglobulin domains in titin (Ig27, Fig. 7) is stabilized by hydrophobic interactions between the sheets. β–barrel is the example of β–class protein, in which a single β–sheet is wrapped in such a way that its edge β–strands form HB interactions. The example of such fold is given in Fig. 8. A new interesting example of β–class is β–helix fold shown in Fig. 9. This type of structure is possibly used by amyloid assemblies.

Fig. 6 Native structure of myoglobin demonstrates the example of folded leaf fold (PDB access code 5mba).

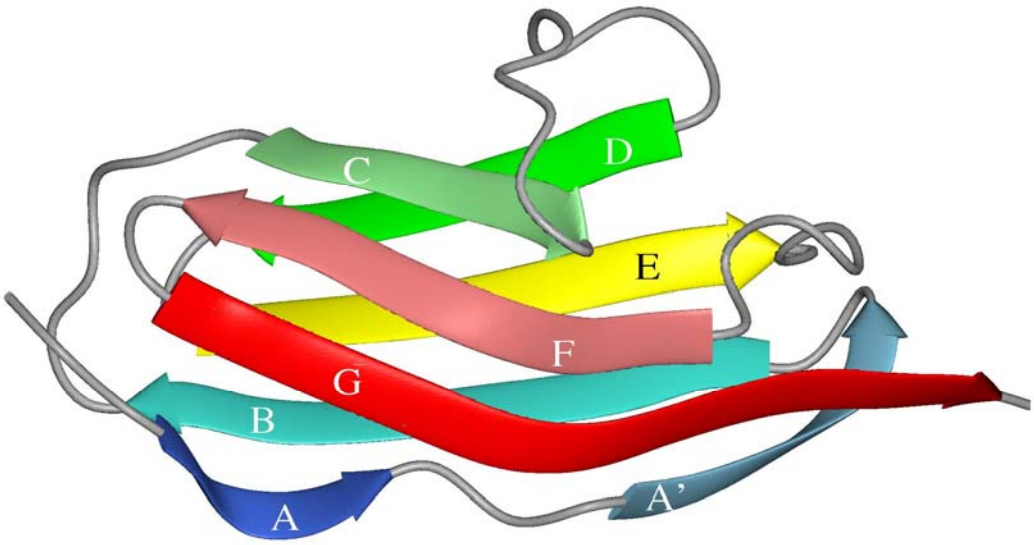

Fig. 7 Native structure of Ig27 domain represents the β-sandwich fold (PDB access code 1tit).

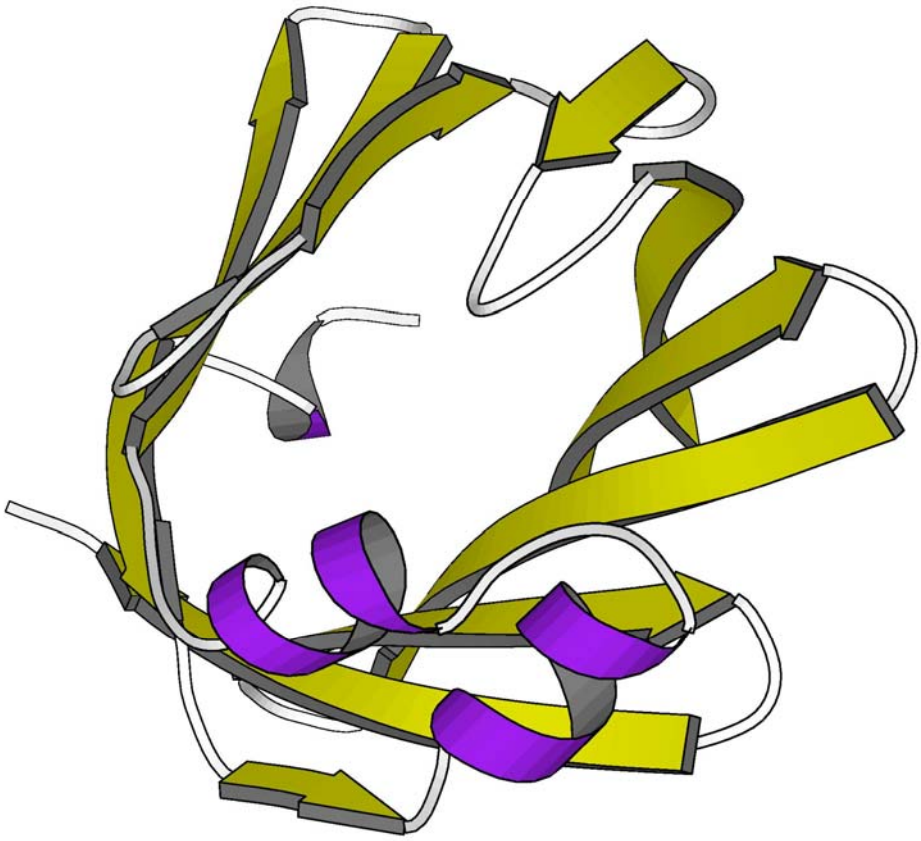

Fig. 8 Native structure of fatty acid binding protein (PDB access code 1hms) forms a β–barrel.

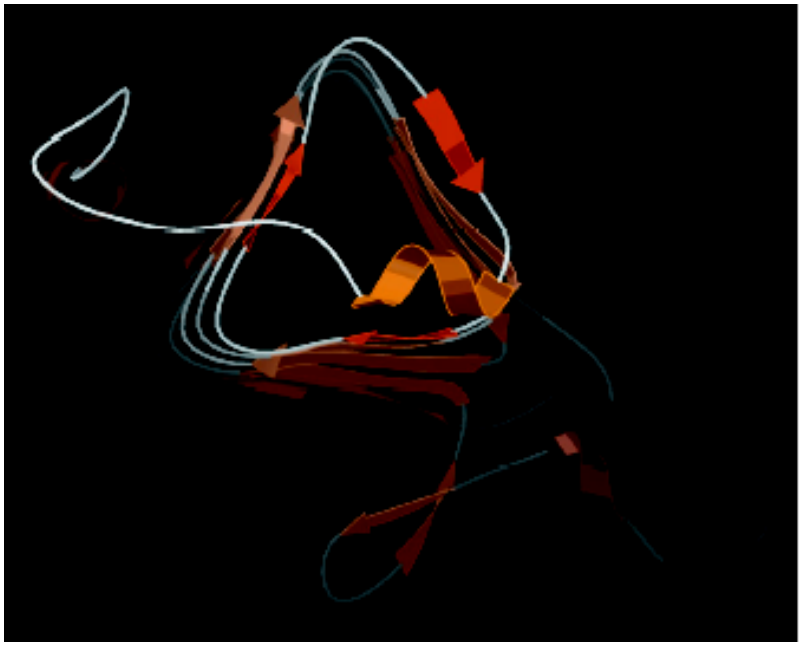

Fig. 9 Native structure of carbonic anhydrase (access code 1qre) takes the form of β-helix fold.

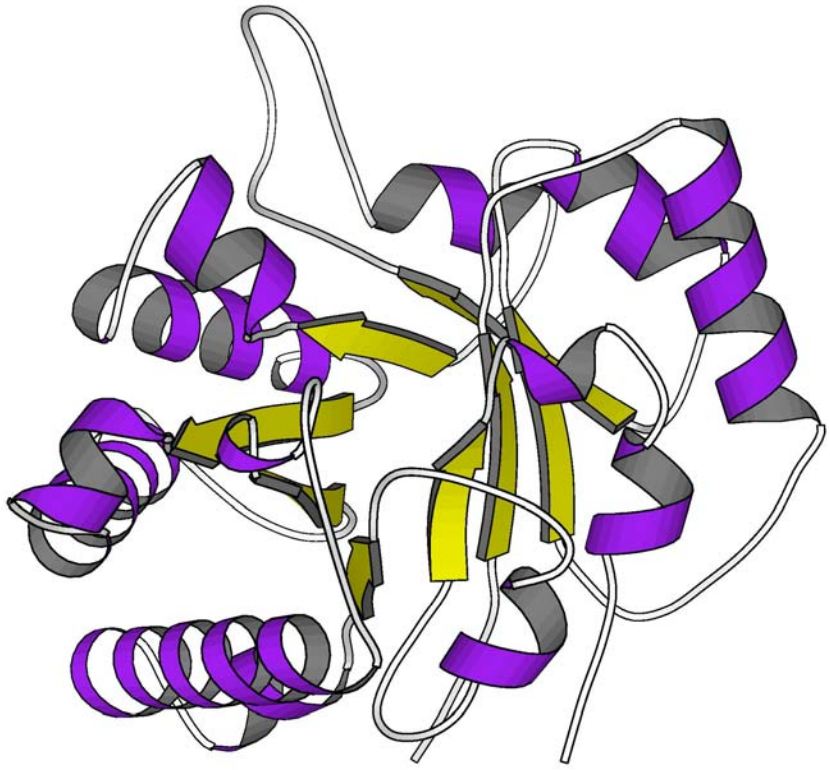

Fig. 10 Native structure of triosephosphate isomerase (PDB access code 1tim) is the example of α/β barrel.

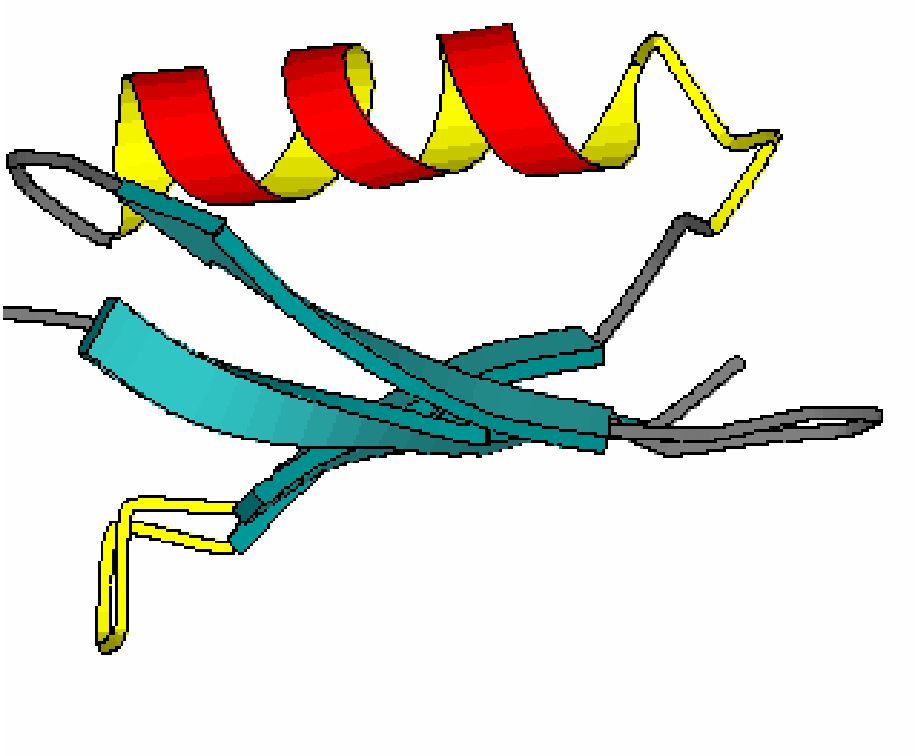

Fig. 11 Native structure of GB1 protein (PDB access code 2gb1) shows the example of one of the folds in the α+β class.

A variety of folds may be distinguished within α/β and α+β classes. For example, the native structure of triosephosphate isomerase represents one of the most common α/β fold, an α/β barrel, in which α–helices and β–strands are intermixed (Fig. 10) and wrapped in the form of a barrel. A simple fold of α+β class is given by the native conformation of the B1 domain of protein G (Fig. 11). In this structure a single α–helix is packed against four stranded β–sheet. Another example of the native structure with clearly separated α and β structure is shown in Fig. 12. Native state of lysozyme has small β–sheet domain, which is positioned on the side of the much larger α–helix domain.

In this lecture only a brief introduction into the structural classification of proteins is presented. For example, the concepts of protein family or superfamily, which are based on sequence comparisons or evolutionary considerations, are beyond the scope of this course. Further information may be found in the papers by Chothia and coworkers (*Methods in Enzymology* **266**, 635 (1996); *Current Opinions in Structural Biology* **7**, 369 (1997)) or at SCOP website.

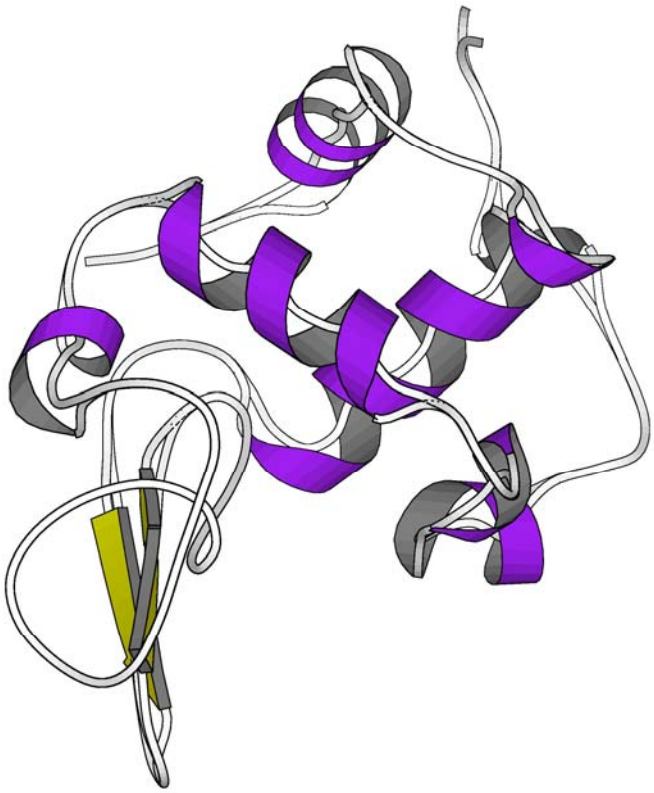

Fig. 12 Native structure of lysozyme (PDB access code 1rey) belongs to α+β class.

VI. Simulation of lattice models probe the number of protein folds

Because of the limited number of ways, which achieve tight packing of secondary structure elements, the total number of folds N_s appears to be restricted as well. It is estimated that N_s is of an order of 10^3 or so (*Current Opinions in Structural Biology* **7**, 369 (1997)). In order to rationalize this suggestion the lattice simulations can be used (*Advances in Chemical Physics* **120**, 35 (2001)). Specifically, protein lattice models are useful to probe the scaling of N_s with N, the number of amino acids. Consider a 3D lattice model (see Lectures 3 and 6 for details) and denote $C(N)$ as the number of available conformations for a given N. If there is no excluded volume interactions, $C(N)=z^N$, where z is the lattice coordination number (for 3D cubic lattice z=6). Assume that (i) protein amino acids cannot occupy the same lattice site more than once (the condition of self-avoidance) and (ii) the interactions between amino acids are heterogeneous. Among all conformations we select for further consideration only those which satisfy two basic characteristics of the protein native folds:

a) putative native structure must be compact (i.e., contains maximum number of contacts possible for a given N);

b) putative native state must be of low energy.

Let us now numerically evaluate the number of minimum energy compact structures $C_{MES}(N)$ satisfying these two conditions. To this end all possible conformations for a given N are explicitly enumerated up to N=18. The plot in Fig. 13 shows that the scaling of $C_{MES}(N)$ is qualitatively different from the scaling of the number of compact structures (based on (a) condition) $C_{CS}(N)$ or the number of self-avoiding conformations $C_{SAW}(N)$. $C_{MES}(N)$ does not grow exponentially as C_{CS} or C_{SAW} do, but instead almost levels off with N. This result suggests that the space of conformations capable of encoding native states of proteins is very sparse and the number of such structures is limited. The calculations presented in Fig. 13 support the idea that all (or almost all) protein folds may be successfully determined.

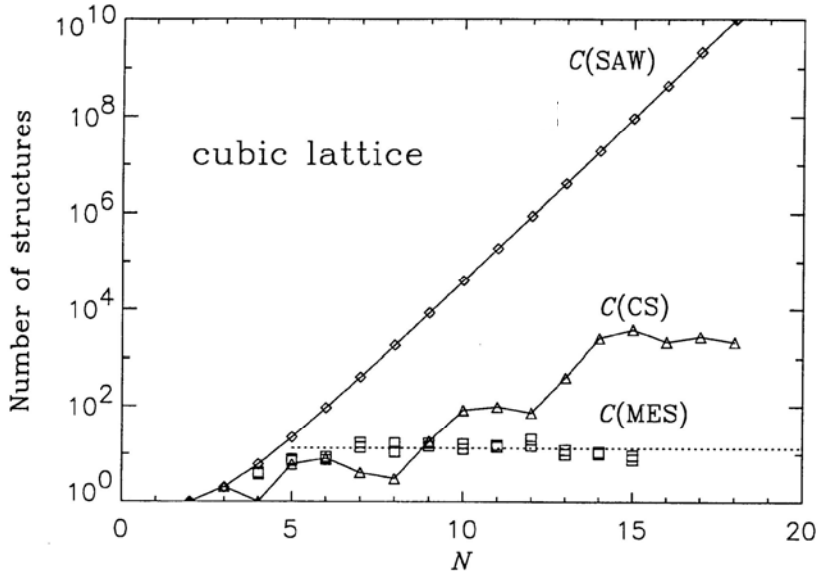

Fig. 13 Scaling of the number of self-avoiding conformations C_{SAW}, the number of compact structures C_{CS} and the number of minimum energy compact structures C_{MES} with the number of amino acids N.

Appendix: Chou-Fasman parameters

Individual residues show remarkably different propensities to accommodate α–helix structure. The simplest computation of amino acid structural preferences was done by Chou and Fasman (*Biochemistry* **13**, 222 (1974)). Consider a database of sequences with known structures. Let the number of residues is N and the number of residues in the α–helix conformation is N_h. The average probability to observe an α–helix for generic residue (or at a random position in the database sequence) is $P_h=N_h/N$. Let us now choose the residue type i and determine the number of residues i in the database $N(i)$. Compute also the number of the residues i, which are found in the α–helix conformation, $N_h(i)$. The probability $P_h(i)=N_h(i)/N(i)$ gives the actual probability to observe residue type i in the α–helix.

To evaluate the propensity of the residue type *i* to form α–helix, consider the ratio $p_h(i)=P_h(i)/P_h$. The actual Chou-Fasman values in the Table 1 are computed as $p_h(i) = ln\ P_h(i)/P_h + 1$. The value of $p_h(i)=1$ indicates that the residue *i* has no structural preference. The values of $p_h(i) < 1$ or > 1show that the residue *i* is either poorly or well accommodated by the helical conformation, respectively. Similarly, the propensity to form β–strand structure is defined as $p_s(i)= ln\ P_s(i)/P_s + 1$ (or turn $p_t(i)$). Chou-Fasman parameters are only meaningful, if sufficient diversity of proteins structure is observed in the database used for their computation.

Table 1. Chou-Fasman parameters for amino acids.

i	ph(i)	ps(i)	pt(i)
Alanine	1.42	.83	.66
Arginine	.98	.93	.95
Aspartic Acid	1.01	.54	1.46
Asparagine	.67	.89	1.56
Cysteine	.70	1.19	1.19
Glutamic Acid	1.51	.37	.74
Glutamine	1.11	1.10	.98
Glycine	.57	.75	1.56
Histidine	1.00	.87	.95
Isoleucine	1.08	1.60	.47
Leucine	1.21	1.30	.59
Lysine	1.14	.74	1.01
Methionine	1.45	1.05	.60
Phenylalanine	1.13	1.38	.60
Proline	.57	.55	1.52
Serine	.77	.75	1.43
Threonine	.83	1.19	.96
Tryptophan	1.08	1.37	.96
Tyrosine	.69	1.47	1.14
Valine	1.06	1.70	.50