Protein primary structure: Amino acids

I. Diversity of proteins

Numerous biological functions are performed by proteins. These include oxygen transport and storage (hemoglobin and myoglobin, respectively), muscle functioning (titin), growth of bones (collagen), cell adhesion (fibronectin) etc. The diversity in functions suggests a wide diversity in protein structures. Indeed, proteins show remarkable variation in size and shape. The largest protein, titin, consists of about 30,000 amino acids and includes approximately 300 domains. Collagen is composed of three individual chains, which are wound in a triple helix elongated structure (Fig. 1). One of the smallest proteins is WW domain (FBP28, shown in Fig. 1), which contains from 35 to 40 amino acids and is stable without disulfide bonds. The WW domains adopt a β-sheet structure in their native state, which includes only three β-strands.

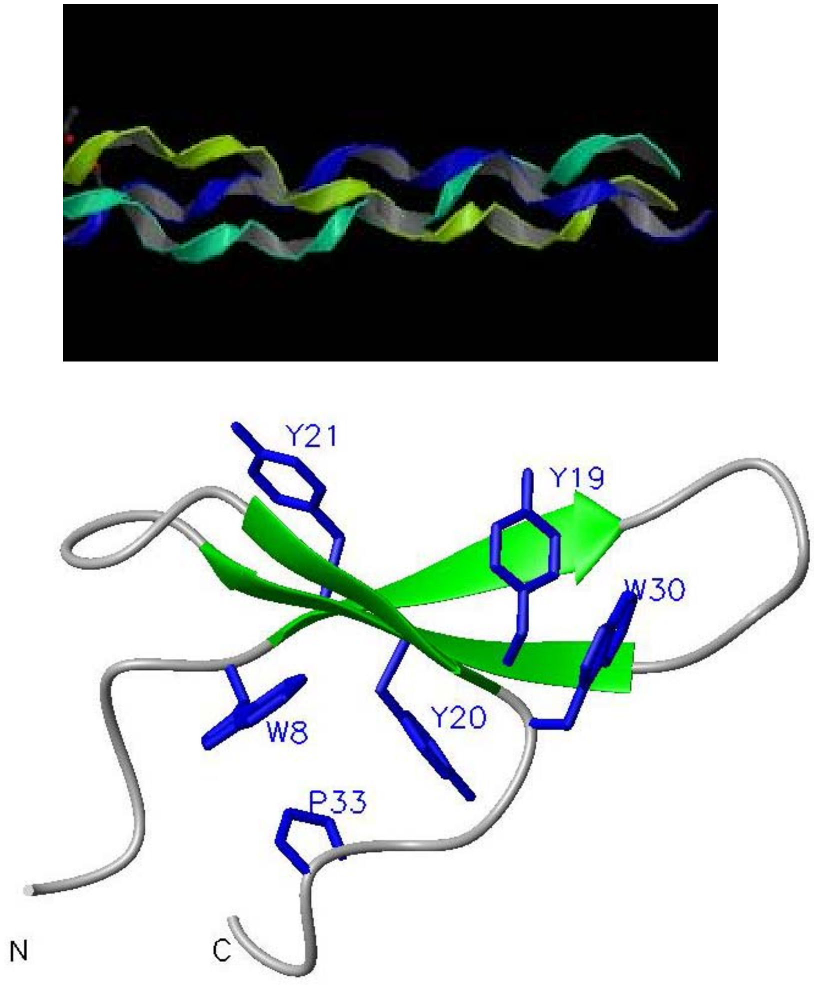


Fig. 1 Native conformations of triple helix collagen (PDB access code 1cag, upper panel) and three β-strand FBP28 WW domain (PDB access code 1eol, lower panel).

The diversity of proteins is also reflected in the existence of three classes of proteins, namely globular, membrane and fibrous proteins. Collagen and WW domains belong to fibrous and globular proteins, respectively.

II. Water structure

Water is an excellent solvent and plays a critical role in determining the structure and stability of proteins. Despite the simplicity of its molecular structure, water shows very unusual properties. For example, water expands upon freezing transition and, in fact, expands even in the liquid form when temperature is reduced from 4°C to 0°C. Water also has an unusually large heat capacity (specific heat). The structure of liquid water is mainly determined by the formation of hydrogen bonds (HB, for definition see below). Each water molecule forms four HBs, two of which are formed by an oxygen atom and another two - by two hydrogen atoms (Fig. 2). Furthermore, because the three-dimensional distribution of HBs is far from being planar, water molecules are arranged in molecular tetrahedrons.

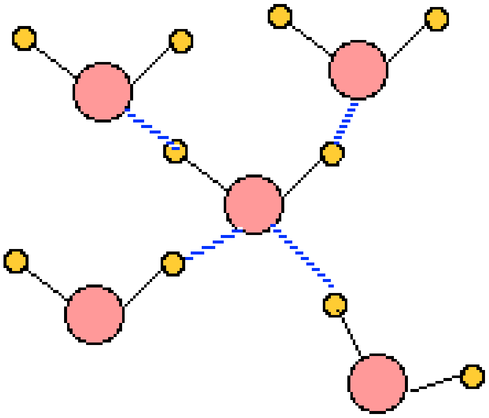


Fig. 2 Schematic representation of liquid water structure. Each water molecule is engaged in four HBs with four other waters shown by blue dotted lines. Oxygen and hydrogen atoms are displayed in pink and yellow, respectively.

Hydrogen bonding D-H…A is a weak electrostatic interactions resulting from sharing of a hydrogen atom H between the donor atom D with the acceptor atom A. The hydrogen atom H and the donor atom A are covalently linked. Both donor and acceptor atoms carry partial negative charges. In proteins, a classical intraprotein HB is formed between backbone amide and carbonyl groups. In this case, D and A are the backbone nitrogen and carbonyl oxygen atoms. The formation of HB can be observed by using radial density distribution functions *g(r)* for the atom pairs H and A or D and A. Specifically, in the case of H..A pair of atoms *g(r)* gives the density of the atoms A at the distance *r* from the hydrogen H. Consider now the structure of water around protein backbone. In this example, H is amide hydrogen and O is water oxygen (Fig. 3). A well defined peak of *g(r)* indicates a local build-up of water density at the distance of about 1.8Å, which is associated with the formation of hydration shell (out of water molecules) around protein backbone. The maximum in Fig. 3 is due to the formation of HB between water oxygens and protein amide hydrogens. For the D and A pair of atoms the maximum in *g(r)* function is reached at about 2.9 Å. The protein-water HBs are also formed between water hydrogens and backbone carbonyl oxygens as well as between water molecules and protein side chains. Protein-protein HBs also include those formed between side chains, between side chains and backbone, and also weak and rare HBs involving C_α-carbons as a donor atom (of the C_α-H…O type).

There is no universal definition of a HB, although two empiric definitions, geometric and energetic, are commonly used. The geometric definition states that a HB is formed, if the distance $r_{DA} \leq 3.5$Å and the angle D-H…A is greater than 120° (note that other numerical values for r_{DA} and DHA angle are sometimes used). The energetic definition applied for intraprotein backbone HBs relies on the computation of electrostatic energy E_{hb} due to the interactions between hydrogen H and nitrogen N of the amide group and oxygen O and carbon C atoms of the carbonyl group. This definition was first proposed by Kabsch and Sanders (Biopolymers **22**, 2577 (1983)) and is used in DSSP database. The HB is formed, if $E_{hb} \leq -0.5$ kcal/mol, where

$$E_{hb} = 332 q_1 q_2 \left(\frac{1}{r_{NO}} + \frac{1}{r_{HC}} - \frac{1}{r_{HO}} - \frac{1}{r_{NC}} \right) \qquad (1)$$

In Eq. 1 r are the distances between respective atoms, q_1=0.2 and q_2=0.42 are the partial charges on (hydrogen, nitrogen) and (carbon, oxygen) pairs, respectively, and 332 is a factor expressed in kcal·mol^{-1}·Å. Partial charges are expressed in the units of e, which is an absolute value of electron charge. The typical energy of a HB is several kcal/mol.

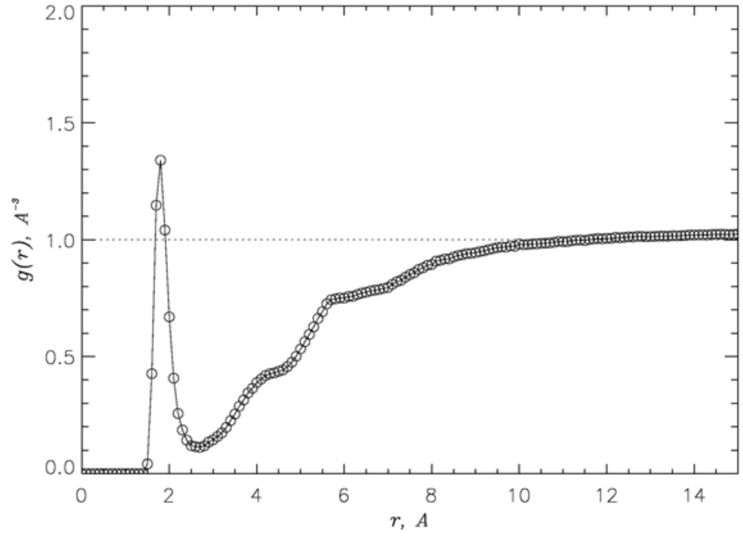


Fig. 3 Radial density distribution function *g(r)* for backbone amide hydrogen H and water oxygen O. The distance *r* represents the separation between H and O.

III. Protein amino acids

The sequence of amino acids forms a primary protein structure. Each amino acid contains central C_α-carbon, hydrogen atom H, protonated NH^+_3, dissociated COO^- group, and side chain R (Fig. 4). This form of amino acid is typical at normal pH. These atomic groups form tetrahedral structure, which leads to two isomer (mirror image) forms (left-handed L-isomer and right-handed D-isomer, see Fig. 5 for the example of alanine isomers). Interestingly, only L-isomers are found in wild-type proteins.

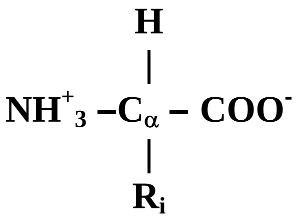


Fig. 4. Structure of amino acid.

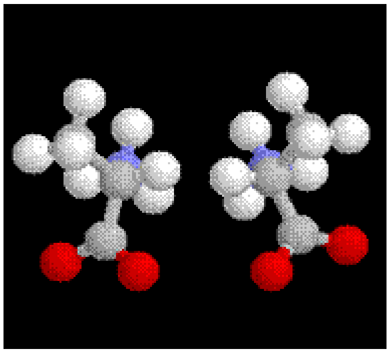


Fig. 5 Spatial tetrahedral structures of L- (left) and D- (right) alanines. Nitrogen and oxygen atoms are shown in blue and red, respectively.

Wild-type proteins, which are synthesized on ribosomes, utilize 20 amino acids. The generic structure of amino acid incorporated in a protein sequence is shown in Fig. 6.

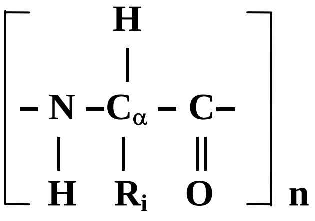
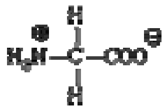


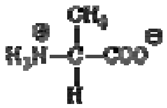
Fig. 6 Structure of generic amino acid *i* in a polypeptide sequence of *n* residues.

In general, amino acids can be divided into hydrophobic, polar (or hydrophilic) and charged residues. Below we consider the structures and properties of individual amino acids.

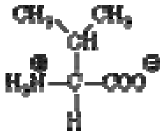
Aliphatic amino acids: These amino acids include glycine, alanine, valine, leucine, and isoleucine. The aliphatic residues are hydrophobic (structures of amino acids below are arranged in the ascending order of their hydrophobicity), have open, sometimes branched side chains. Aliphatic amino acids do not usually form HB and are neutral and not polar. The smallest is Gly, which has no side chain and demonstrates the largest backbone flexibility. As the length of side chain increases, so does the number of available side chain conformations (the *rotamer* library, see below). As the side chains become bulkier, the flexibility of the backbone decreases.



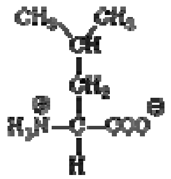
Glycine (Gly, G)



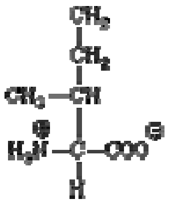
Alanine (Ala, A)



Valine (Val, V)

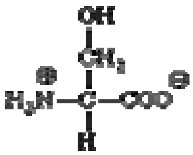


Leucine (Leu, L)

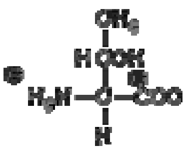


Isoleucine (Ile, I)

Aliphatic hydroxyl amino acids: Serine and threonine belong to this group and contain hydroxyl group OH in their side chains. As a result these amino acids are polar, capable of forming HBs with water and, consequently, are hydrophilic. The side chains of these amino acids are neutral.

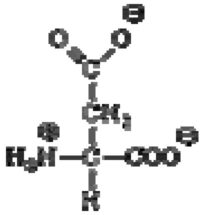


Serine (Ser, S)

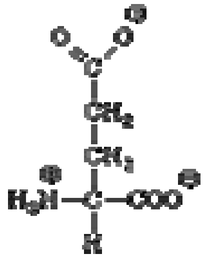


Threonine (Thr, T)

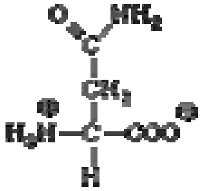
Acidic amino acids and amide derivatives: Asparagine, glutamine, aspartic acid, and glutamic acid are included in this group. Asparagine and glutamine contain amide groups in their side chains and are uncharged. Aspartic and glutamic acids contain deprotonated hydroxyl groups and are negatively charged at pH values above approximately 4. All these four residues are considered polar (and hydrophilic) and capable of forming HBs. Note also that parts of Asp and Glu side chains are hydrophobic.



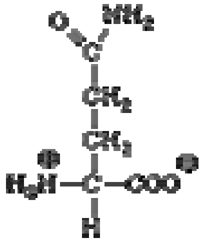
Aspartic acid (Asp, D)



Glutamic acid (Glu, E)

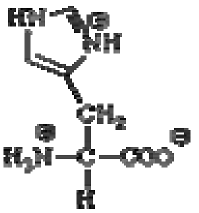


Asparagine (Asn, N)

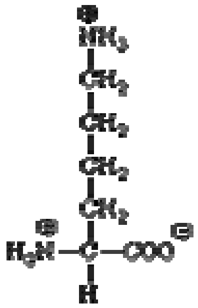


Glutamine (Gln, Q)

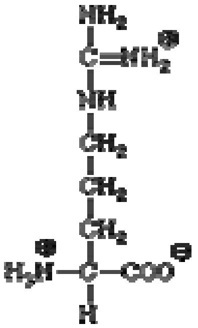
Basic amino acids: The amino acids of basic group are lysine, arginine, and histidine. Lys and Arg contain basic amino groups and are positively charged (protonated) at normal pH. Their pK values (i.e., the values of pH, at which their side chains become neutral) are in the range of 10 to 12. His side chain has the pK value of 6.5, therefore, its charge state depends sensitively on the specific cellular environment. Basic amino acids are highly polar and can participate in hydrogen bonding.



Histidine (His, H)

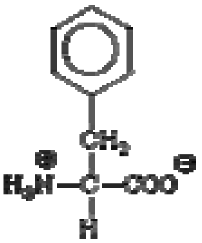


Lysine (Lys, K)

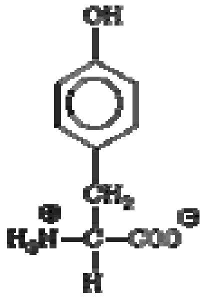


Arginine (Arg, R)

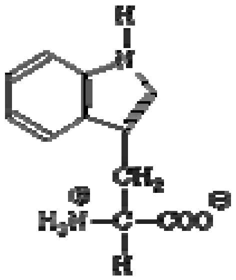
Aromatic amino acids: Phenylalanine, tyrosine, tryptophan amino acids belong to this group. Because the side chains contain aromatic rings, these amino acids are generally hydrophobic, although the degree of their hydrophobicity varies. Phe is the most hydrophobic, while Tyr and Trp are mildly hydrophobic because of partially polar properties of their side chains. These amino acids are uncharged at normal pH. Tyr and Trp, but not Phe, may form HBs



Phenylalanine (Phe, F)

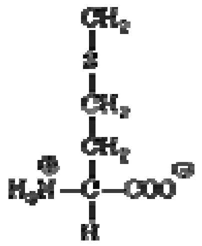


Tyrosine (Tyr, Y)

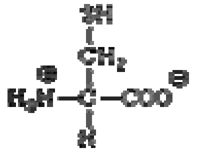


Tryptophan (Trp, W)

Sulfur containing amino acids: Two amino acids, metheonine and cysteine, contain sulfur atoms in the side chains. Met and Cys are hydrophobic and uncharged. As a result of oxidation of SH groups Cys residues can form disulfide bonds, the strength of which are comparable with covalent ones. Their role in folding of BPTI protein was discussed previously.

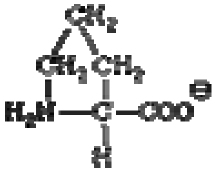


Methionine (Met, M)



Cysteine (Cys, C)

Cyclic amino acids: Proline is the only cyclic amino acid, in which the side chain makes a covalent bond with the amide backbone group. Generally, Pro has aliphatic properties, but because of cyclic side chain its conformational flexibility is highly limited.



Proline (Pro, P)

The frequency of occurrence of amino acids in globular proteins may be evaluated by examining a non-homologous (a maximum sequence similarity of 40%) set of PDB proteins (*Journal of Molecular Biology* **273**, 349 (1997)). In all, the PDB40 dataset contains 971 proteins. The results shown in Table 1 suggests that, in general, frequency of amino acid occurrence ranges from 1 to about 9 %. Two hydrophobic amino acids, Leu and Ala, have the highest frequency of occurrence (>8%), whereas sulfur containing residues, Cys and Met, aromatic amino acid Trp and His appear relatively rare (<3%). It should be noted that the frequencies of amino acids for membrane proteins are skewed towards hydrophobic residues, because these proteins are embedded in lipid bilayers. Similarly, few amino acids, such as Gly or Pro, are greatly overrepresented in fibrous proteins.

Table 1. Frequencies (in %) of amino acids in PDB40 protein dataset

Alanine	8.44
Arginine	4.72
Asparagine	5.98
Aspartic acid	4.68
Cysteine	1.69
Glutamine	3.70
Glutamic acid	6.16
Glycine	7.95
Histidine	2.19
Isoleucine	5.49
Leucine	8.28
Lysine	5.80
Methionine	2.18
Phenylalanine	4.03
Proline	4.61
Serine	6.08
Threonine	5.87
Tryptophan	1.47
Tyrosine	3.61
Valine	6.94

IV. Conformational flexibility in proteins

The flexibility of protein backbone is determined by dihedral angles. Fig. 7 introduces one of the possible definitions of the dihedral angle. Consider four backbone atoms *i,i+1,i+2*, and *i+3* with the radius vectors $\vec{r}_i$, $\vec{r}_{i+1}$, $\vec{r}_{i+2}$, $\vec{r}_{i+3}$. The bond vectors are $\vec{r}_{i,i+1} = \vec{r}_{i+1} - \vec{r}_i$, $\vec{r}_{i+1,i+2} = \vec{r}_{i+2} - \vec{r}_{i+1}$, $\vec{r}_{i+2,i+3} = \vec{r}_{i+3} - \vec{r}_{i-2}$. Let us define the vector normal to the plane (*i,i+1,i+2*) as a vector product $\vec{n}_1 = \vec{r}_{i,i+1} \times \vec{r}_{i+1,i+2}$. Similarly, we define the vector normal to (*i+1,i+2,i+3*) plane $\vec{n}_2 = \vec{r}_{i+1,i+2} \times \vec{r}_{i+2,i+3}$. The angle between the vectors $\vec{n}_1$ and $\vec{n}_2$ is called the dihedral angle ϕ, which may be computed as $\phi = \cos^{-1}\left(\vec{n}_1\vec{n}_2 / \left(\left|\vec{n}_1\right|\left|\vec{n}_2\right|\right)\right)$. The sign of ϕ is determined by the sign of the scalar product ($\vec{n}_1\vec{r}_{i+2,i+3}$). Negative sign of this scalar product corresponds to ϕ<0. The dihedral angle may be also defined as $\phi = \cos^{-1}\left(\vec{t}_1\vec{t}_2 / \left(\left|\vec{t}_1\right|\left|\vec{t}_2\right|\right)\right)$, where $\vec{t}_1$ and $\vec{t}_2$ are the vectors normal to the line *(i+1,i+2)* and lying in the planes (*i,i+1,i+2*) and (*i+1,i+2,i+3*), respectively. In this case the sign of ϕ is determined as $sign(\phi) = sign\left((\vec{t}_1 \times \vec{t}_2)\vec{r}_{i+2,i+3}\right)$. The *trans* conformation of four atoms corresponds to ϕ=180°, *cis* state is represented by ϕ=0°, and *gauch* states are ϕ=±60°.

Two dihedral angles ϕ and ψ defined for the protein backbone (see Lecture 3, Fig. 1) determine completely the conformational flexibility of the backbone. The third dihedral angle ω defined by the four atoms $C_{\alpha i}$,C_i,N_i,$C_{\alpha,i+1}$ of the residues *i* and *i*+1 is frozen in the *trans* state due to the double bond character of the peptide bond. This arrangement sets C_α-carbons at the maximum spatial separation. Steric clashes between neighboring side chains severely restrict the values of accessible ϕ and ψ as it is seen from Ramachandran plots.

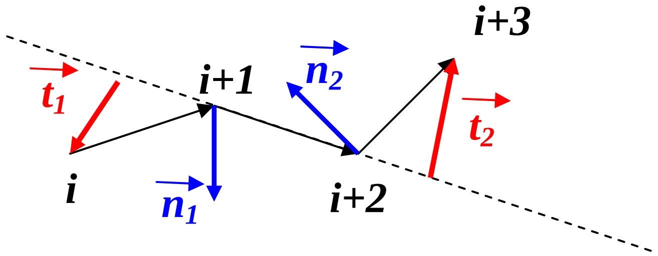


Fig. 7 Backbone vectors used in the definition of dihedral angle ϕ.

In a similar way, the dihedral angles may be defined for the long side chains. The side chain dihedral angles χ_i are allocated as

$$\mathbf{C_\alpha - \underset{\chi_1}{CH_2} - \underset{\chi_2}{CH_2} - \underset{\chi_3}{CH_2} - \ldots}$$