CPS260/BGT204.1 Algorithms in Computational Biology **November 04, 2003**

Lecture 19: Proteins, Primary Struture

Lecturer: Pankaj K. Agarwal *Scribe: Qiuhua Liu*

19.1 The Building Blocks of Protein [1]

Proteins are polypeptide chains obtained by translation from strands of messenger RNA. The functional properties of proteins depend on their three-dimensional structures. The three-dimensional structure arises because particular sequences of amino acids in polypeptide chains fold to generate, from linear chains, compact domains with specific three-dimensional structures(Figure 19.1). Totally there are 20 different amino acids. The amino acid sequence of a protein's polypeptide is called its **primary** structure. Different regions of the sequence form local regular **secondary** structures, such as alpha (α) helices or beta (β)strands. The **tertiary** structure is formed by packing such structural elements into one or several compact globular units called domains. The final protein may contain several polypeptide chains arranged in a **quaternary** struture. By formation of such tertiary and quaternary structure amino acids far apart in the sequence are brought close together in three dimensions to form a functional region.

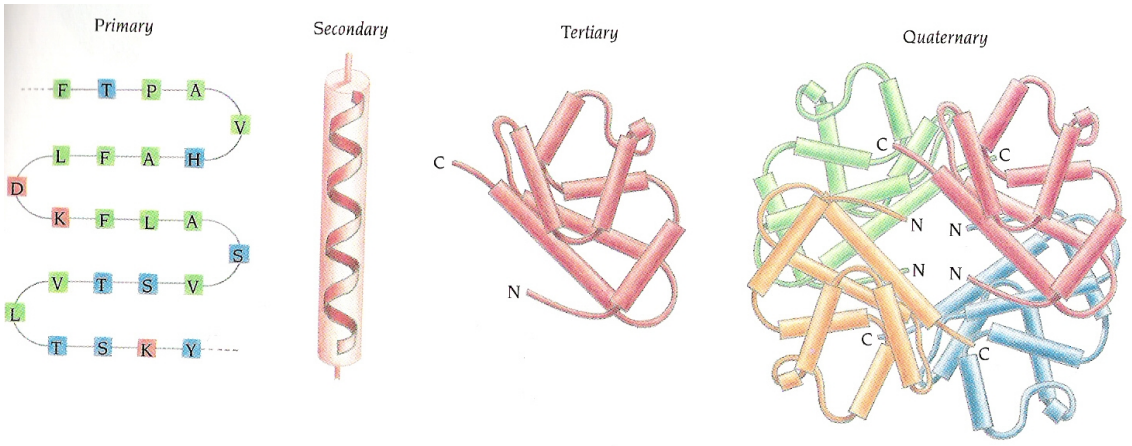

Figure 19.1: The protein structures [1]

To understand the biological function of proteins we would like to be able to deduce or predict the three-dimensional structure from the amino acid sequence. However, this folding problem is still unsolved and remains one of the most basic intellectual challenges in molecular biology. Instead, the three-dimensional structures of individual proteins are determined experimentally by x-ray crystallography, electron crystallography or neclear magnetic resonance (NMR) techniques.

19.1.1 Proteins are polypeptide chains

All of the 20 amino acids have in common a central carbon atom (C_α) to which are attached a hygrogen atom, an amino group (NH_2), and a carboxyl group (COOH) (Figure 19.2a). What distinguishes one amino acid from another is the side chain attached to the C_α through its fourth valence. Amino acids are joined end-to-end during protein synthesis by the formation of **peptide bonds** which takes place when the carboxyl group of one amino acid condenses with the amino group of the next to eliminate water (Figure 19.2b). This process gets repeated as the chain elongates. The resulting repeating sequence of nitrogen, α-carbon and carbon atoms is the **backbone** or **main chain** of the protein. Amino acids that are linked into the polypeptide are referred to as **residues**.

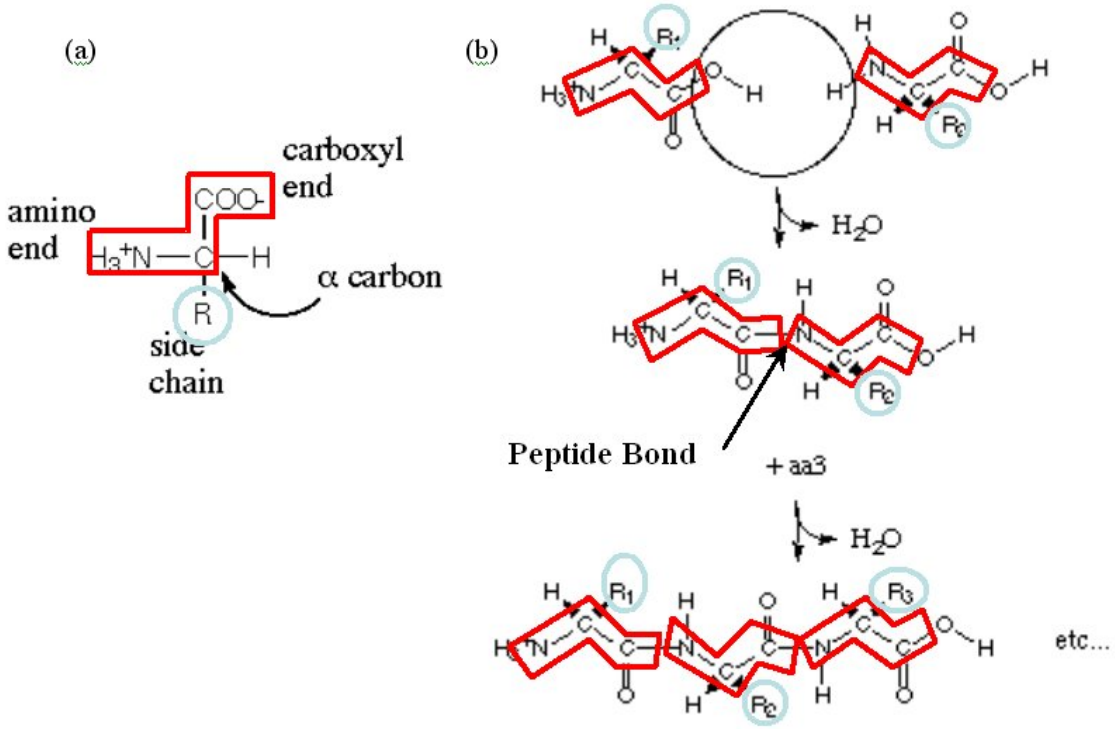

Figure 19.2: Proteins are built up by amino acids that are linked by peptide bonds. (a) Schematic diagram of an amino acids. (b) The amino acid residues are joined together by a peptide bond [2].

The four neighbours of an α-carbon, C_α, are at the vertex positions of a tetrahedron around C_α. This tetrahedron has two orientations, one being the mirror image of the other, as illustrated in Figure 19.3. The two oriented forms are referred to as **isomers** and distinguished by letters L and D. Only L-amino acids occur in nature as building blocks of proteins.

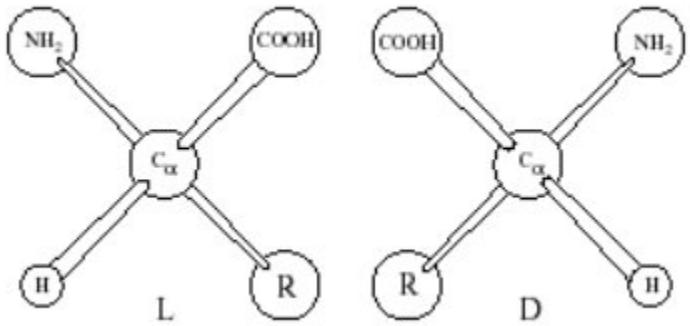

Figure 19.3: The two isomers of an amino acids [3].

19.1.2 The genetic code specifies 20 different amino acids

As mentioned earlier, there are 20 different amino acid side chains in all, specified by the genetic code. The sequence of nucleotides is read in groups of three, called codons. Totally we have $4^3 = 64$ codons (Figure 19.4). The 20 amino acids are usually divided into three different groups defined by the chemical nature of the side chain (Figure 19.5): hydrophobic, hydrophilic and in-between. Their names are abbreviated with a three-letter code.

19.1.3 Ramachandran Plot

Since the peptide units are effectively rigid groups that are linked into a chain by covalent bonds at the C_α atoms, the only degrees of freedom they have are rotations around these bonds. Each unit can rotate around two such bonds: the C_α- C' and the N-C_α bonds. By convention, the angle of rotation around the N-C_α bond is called **phi** (ϕ) and the angle around the C_α-C' bond from the same C_α atoms is called **psi** (ψ). In this way, the conformation of the whole main chain is completely determined when the ϕ and ψ angles for each amino acids are defined.

Most combinations of ϕ and ψ angles of an amino acids are not allowed because of steric collisions between the side chains and main chain. The angle pairs ϕ and ψ are usually plotted against each other in a diagram called a **Ramachandran plot** after the indian biophysicist G.N.Ramachandran who first made calculations of sterically allowed regions. Figure 19.6 shows the results of such calculation for all amino acids except glycline from a number of accurately determined protein structures. The major allowed regions in Figure 19.6 are the right-handed α -helical cluster (Figure 19.7) in the lower left quadrant; the broad region of extended β strands (Figure 19.7) of both parallel and antiparallel β structures in the upper left quadrant; and the small, sparsely popluated left-handed α-helical region in the upper right quadrant.

SECOND POSITION

FIRST POSITION	U	C	A	G	THIRD POSITION
U	phenyl-alanine	serine	tyrosine	cysteine	U
					C
	leucine		stop	stop	A
			stop	tryptophan	G
C	leucine	proline	histidine	arginine	U
					C
			glutamine		A
					G
A	isoleucine	threonine	asparagine	serine	U
					C
			lysine	arginine	A
	* methionine				G
G	valine	alanine	aspartic acid	glycine	U
					C
			glutamic acid		A
					G

Figure 19.4: The genetic code [2].

19.1.4 Protein Structure Classification-CATH [5]

The CATH is a protein structure database which curretly contains more than 1200 evolutionary superfamilies, constructed by both automatic and manual evaluation of structure relationships. The first level in the CATH hierachy describes the protein (C)lass; that is whether the strcture comprises mainly α-helices, mainly β-strands or a mixture of both. At the next (A)rchitectural level, proteins are grouped according to the orienetations of their secondary structures in 3-D. A large portion of structures adopt very simple layered architectures such as sandwiches (e.g. two or three-layer α-β proteins) or barrel-like arrangements (Figure 19.8). The (T)opology level or fold group then discriminates according to differences in the connectivities between the secondary structures in these architectures.

19.2 Representation of the Backbones for Proteins

As mentioned in the first section, the sequence of C_α atoms of a protein is called its backbone. The representation of the backbones are listed below.

- Coordinates of C_α atoms

  If a protein has n amino acids, it will need $3n$ real numbers $(x_1, y_1, z_1, x_2, y_2, z_2, ..., x_n, y_n, z_n)$ to

represent its backbone. The problem is that usually the proteins contain thousands of amino acids. Therefore, we need to compress the presentation.

- Coordinates of the first C_α atom and (ϕ, ψ) angles for the rest

  If a protein has n amino acids, this representation will need $2n + 1$ real numbers $(x_1, y_1, z_1, \phi_2, \psi_2, ..., \phi_n, \psi_n)$. This method still needs a lot of space.

- HP model

  In the HP model [6], a protein is modelled as a sequence of hydrophobic (H) and hydrophilic (P) monomers. The sequence is grown into a two-dimensional lattice using a self-avoiding walk, and the resulting conformation is calculated by summing interactions between pairs of monomers that occupy adjacent lattice sites but are not covalently bounded (Figure 19.9). Each grid point has 4 neighbours, and for each of the C_α atoms, 2 of its neighbours are occupied by its adjacent C_α atoms.

19.3 Protein Structure Alignment

19.3.1 Structure similarity measure

Many protein structure alignment algorithms use geometry for comparison purposes, but ignore the similarities in the enviroment of the residues. To account for the structure similarity, two different root mean square (RMS) values have been proposed, $cRMS$ and $dRMS$. Given an alignment of proteins A and B (Figure 19.10), where the dashed lines represent the alignment between the two corresponding C_α atoms i_r and j_r in A and B, $r = 1, ...k$), the $cRMS$ is defined as the norm of the distance vector of the alignment:

$$cRMS = \sqrt{\frac{1}{k}\sum_{r=1}^{k} d^2(A(i_r) - B(j_r))}, \tag{19.1}$$

where k is the number of atoms that are aligned, $A(i_r)$ and $B(j_r)$ are the transformation (including translation and rotation) of the atoms indexed i_r and j_r and $d(.,.)$ represents the Euclidean distance.

The $dRMS$ measures the difference between the respecitve distance matrices of the alignment:

$$dRMS = \sqrt{\frac{1}{\binom{k}{2}} \sum_{1 \le r < s \le k} |d(A(i_r) - A(i_s)) - d(B(j_r) - B(j_s))|^2}, \tag{19.2}$$

19.3.2 Structure alignment algorithm

The alignment algorithm using $cRMS$ and $dRMS$ including two steps:

- Given an alignment, define the score of the alignment by $cRMS$ and $dRMS$ as,

$$\sigma(A, B) = \frac{1}{1 + cRMS} + G \tag{19.3}$$

and

$$\sigma(A, B) = \frac{1}{1 + dRMS} + G \tag{19.4}$$

where G represents the gap penalty.

- Find an alignment that maximizes the score:

  -Translate and Rotate B

  -For a fixed embedding of B, compute an alignment that maximize the score

  The above two steps can be done by dynamic programming.

For $dRMS$, the score is not affacted by translation and rotation, therefore, in the second step we only need to find an alignment that maximizes the score (Equation 19.3);

For $cRMS$, in the second step we need to find the transformation (translation and rotation), such that an aligment minimizes the score, which is done by the EM alogrithm:

- Fix an initial embedding of B;
- Find an alignment μ that maximizes the score (Equation 19.4);
- Find the translation and rotation for μ that maximize the score (Equation 19.4).

Like all the EM algorithms, this iteration method can get trapped in a local optima. To reduce the probability of this, one usually tries multiple initial embeddings of B or appplies the simulated annealing method.

19.3.3 Programs of structure alignment

Many research groups in structure alignment have generously made their programs available for use over the Internet and the World Wide Web. Here, we give four popularly used alignment algorithms and their websites:

- DALI: *http://www2.ebi.ac.uk/dali*
- STRUCTAL: *http://bioinfo.mbb.yale.edu/align/server.cgi*.
- LOCK: *http://gene.stanford.edu/lock/* .
- VAST: *http://www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html*.

Among them, DALI uses $dRMS$ as the similarity measure and both STRUCTAL and LOCK use $cRMS$ as the similarity measure. VAST is based on graph heuristic appoarch and are different from the other three. For more information about those programs, please refer to their websites.

References

[1] C. Branden and J. Tooze, “Introduction to Protein Structure”, Second Edition, Chapter 1,1999.

[2] P.K. Agarwal, class powerpoint notes: Protein.pdf.

[3] *http://www.cs.duke.edu/education/courses/fall02/cps296.1/*

[4] *http://www.expasy.org/swissmod/course/text/chapter1.htm*.

[5] C.A. Orengo and et al., “The CATH protein family databases: A resource for structural and functional anotation of genomes”, Proteomics 2002, 2:11-21.

[6] P. Keohl, “Protein structure similarities”, Current Opinion in Structure Biology 2001, 11:348-353.

[7] A. R. Leach, “Molecular Modelling, Principles and Applications”, Second Edition, Chapter 10, 2001.

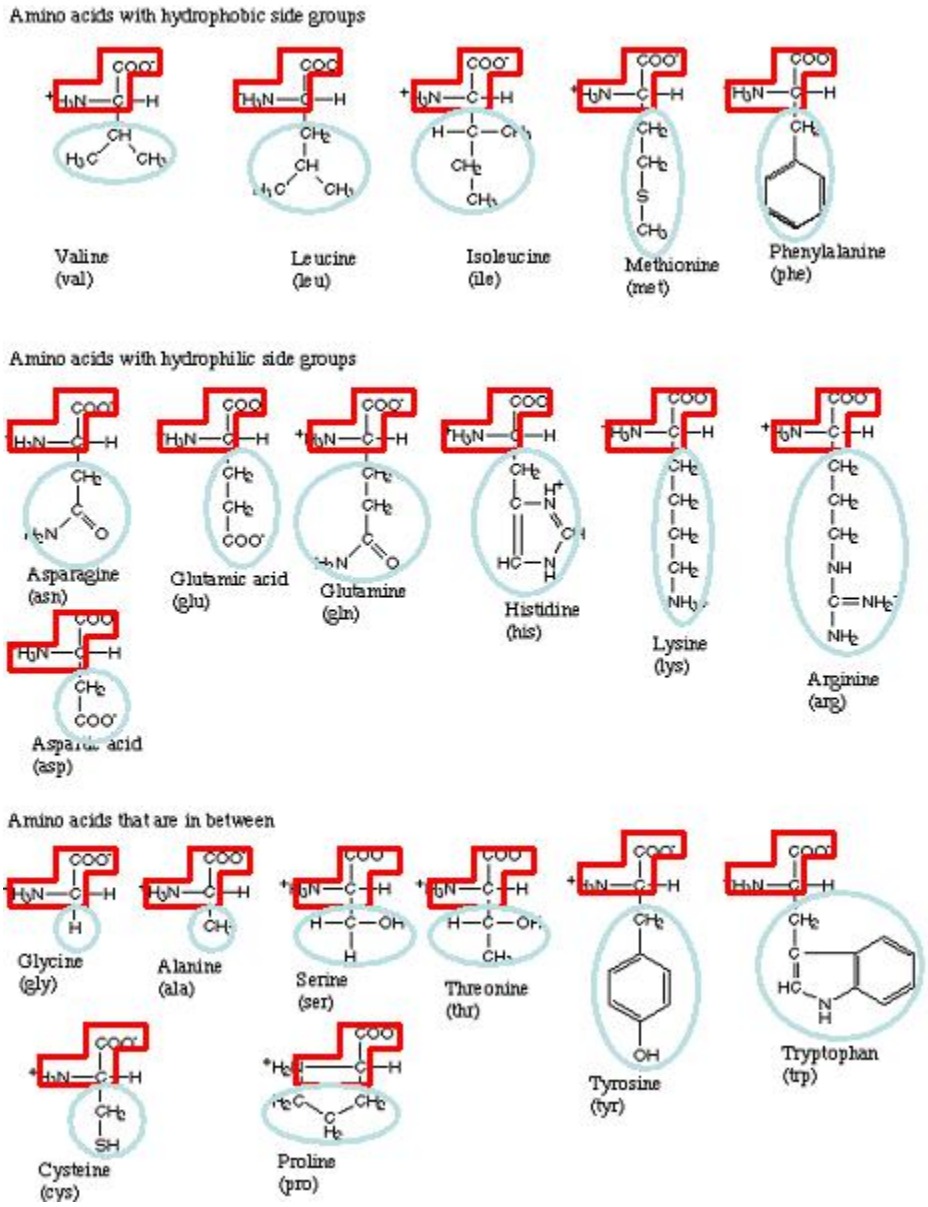

Figure 19.5: The 20 different amino acids [2].

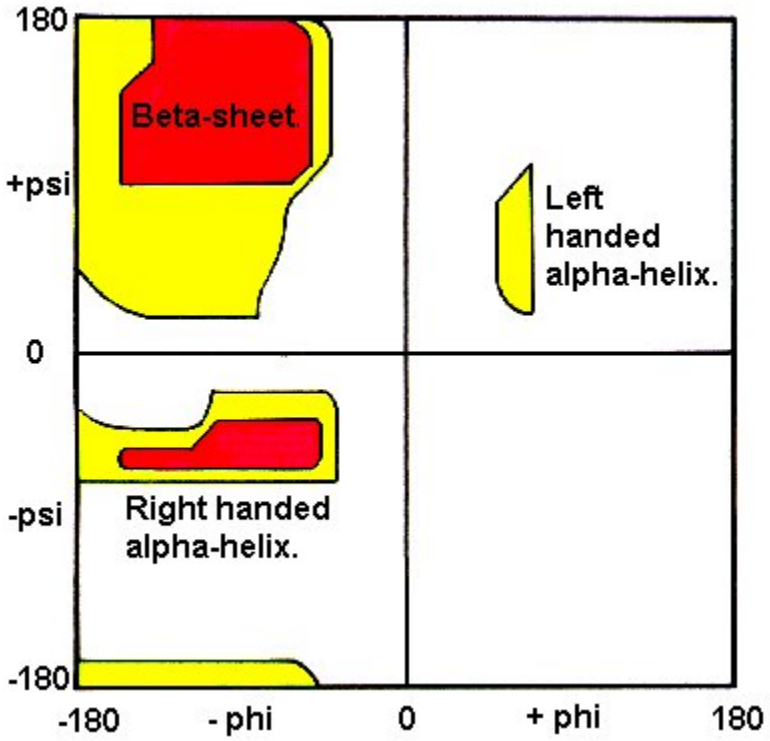

Figure 19.6: Ramachandran plot[4].

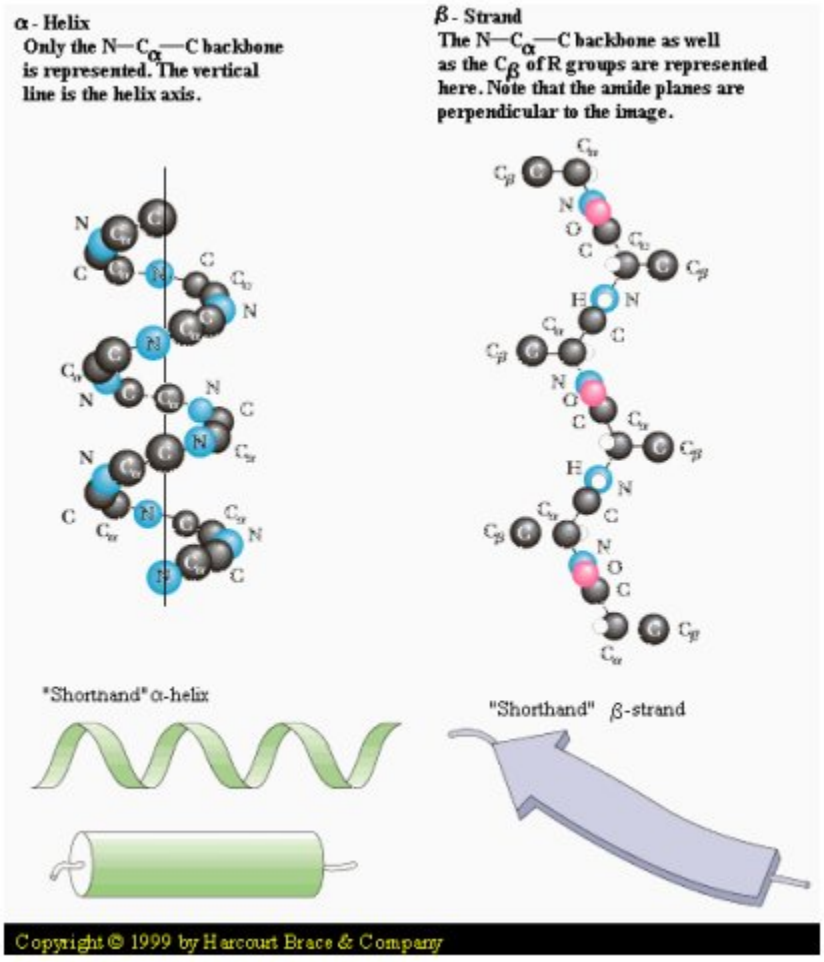

Figure 19.7: α-helix and β-strand [2].

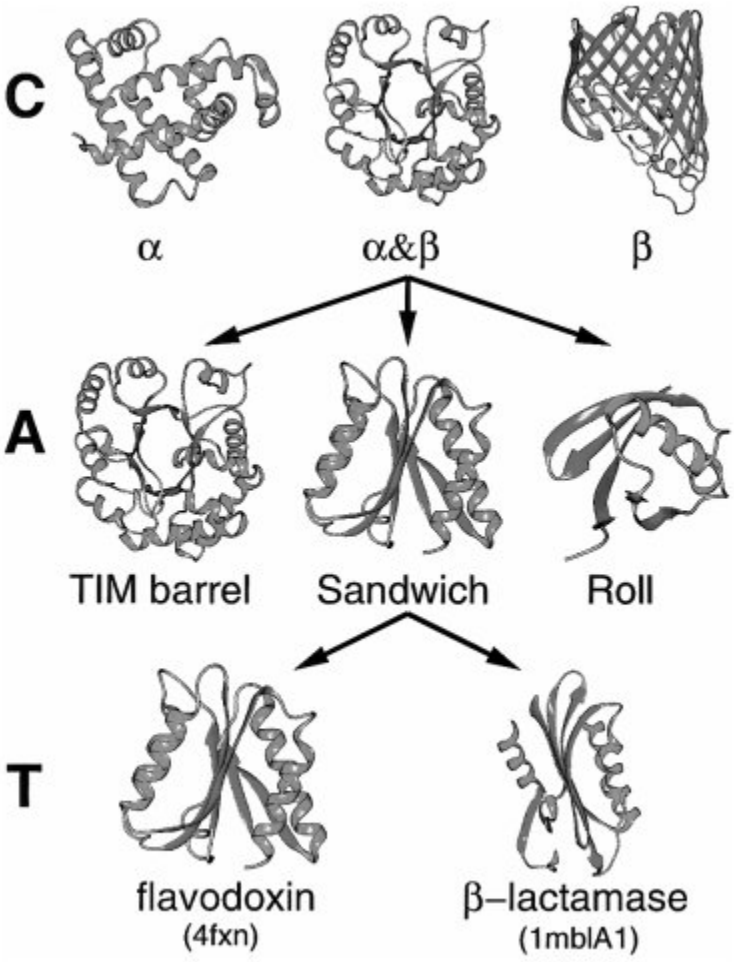

Figure 19.8: Schematic representation of the (C)lass, (A)rchitecture and (T)oplogy levels in the CATH database [5].

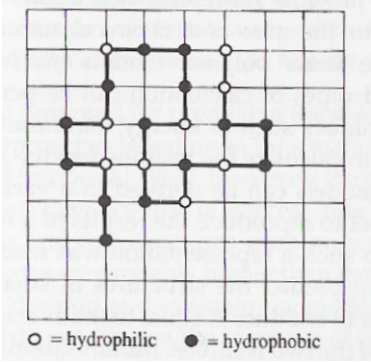

Figure 19.9: The HP model for Backbond Representation [6]

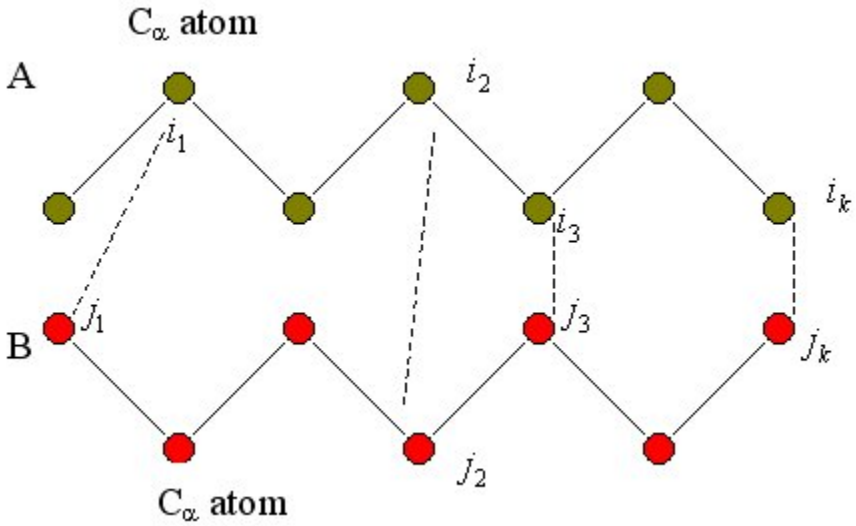

Figure 19.10: Protein structure alignment