Ain Shams University
Faculty of Science

Enhancement of Protein Design Using Electronic Circuits Based on Chou-Fasman Algorithm

A thesis submitted in partial fulfillment of the requirement for the degree of Master of Science
In Biophysics *(2015)*

By
Ibrahim Mahmoud Awaad Darwish

B.Sc. (Biophysics) (2003)
Ain Shams University

Supervised by

***Prof .Dr.* El-Sayed Mahmoud El-Sayed**
Prof. of Biophysics
Physics Department
Faculty of Science
Ain Shams University.

***Prof. Dr.* Salah Yaseen El-Bakry**
Prof. of Theoretical Physics
Head of Physics Department
Faculty of Science
Ain Shams University.

***Prof. Dr.* Amr Khairat Radi**
Prof. Dr. Computational Physics
Physics Department
Faculty of Science
Ain Shams University.

Chapter two

Chapter 2

Theoretical Aspects for Protein Structure

2.1 Protein Structure and Terminology

Proteins are chains of amino acids joined together by peptide bonds. Many conformations of this chain are possible due to the rotation of the chain about each Cα atom. It is these informational changes that are responsible for differences in the 3D structure of proteins. Each amino acid in the chain is polar, i.e. it has separated positive and negative charged regions with a free C=O group, which can act as hydrogen bond acceptor and an NH group, which can act as hydrogen bond donor. These groups can therefore interact in the protein structure. The 20 amino acids can be classified according to the chemistry of the side chain which also plays an important structural role. Glycine takes on a special position, as it has the smallest side chain, only one Hydrogen atom, and therefore can increase the local flexibility in the protein structure. Cysteine on the other hand can react with another cysteine residue and thereby form a cross link stabilizing the whole structure.

The protein structure can be considered as a sequence of secondary structure elements, such as α-helices and β-sheets, which together constitute the overall 3D configuration of the protein chain. In these secondary structures regular patterns of H bonds are formed between neighboring amino acids, and the amino acids have similar Φ and Ψ angles.

The formation of these structures neutralizes the polar groups on each amino acid. The secondary structures are tightly packed in the protein core in a hydrophobic environment. Each amino acid side group has a limited volume to occupy and a limited number of possible interactions with other nearby side chains, a situation that must be taken into account in molecular modeling and alignments. [55]

2.2 Basic Structure of Amino Acids

Amino acids are compounds which have a carboxyl group at one end and an amino group at the carbon atom next to the carboxyl group, the so called α-carbon, Figure (2.1). Several amino acids contain additional acidic or basic groups.

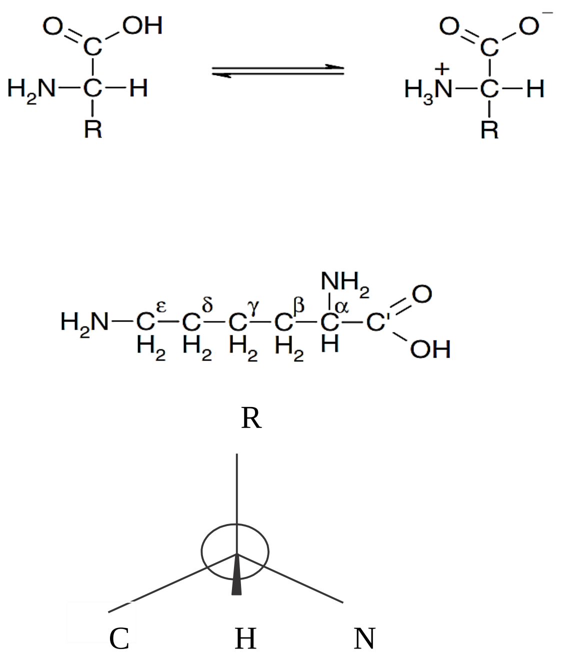

Fig. (2.1). Top: Basic structure of an amino acid. Amino acids can form zwitter-ions. Middle: Nomenclature of carbon atoms, using lysine as example. The Carboxy-carbon is designated C', the following carbon atoms are labeled with the letters of the Greek alphabet. Sometimes the last C-atom is called *ω*, irrespective of the chain length. Bottom: In l-amino acids if the *α*-carbon is placed on the paper plane, with the hydrogen facing you, the remaining substituents read "CORN".

The carboxyl group will donate a proton to the amino group, so that an amino acid (in the absence of other acids or bases) will carry both a negative and a positive charge, making the whole molecule appear uncharged (zwitter-ion).

The simplest amino acid is glycine, where R is a hydrogen atom. Since the α-carbon carries only 3 different ligands (carboxyl group, amino group and hydrogen), it is not enantiomeric. Thus, glycine is not chiral, unlike all other amino acids which carry 4 different ligands on the α-carbon.

2.3 The Isoelectric Point

We know how to determine the pKa of an acid or base, the pH at which half of the molecules are charged. A compound which can act as both acid and base (like an amino acid) has another important property. The isoelectric point *pI*, which is the *pH,* at which the number of positive charges on the molecule is the same as the number of negative charges. At the *pI*, the molecule would therefore appear uncharged. At this *pH,* the molecules ability to interact with water is the lowest, and therefore its solubility is the lowest.

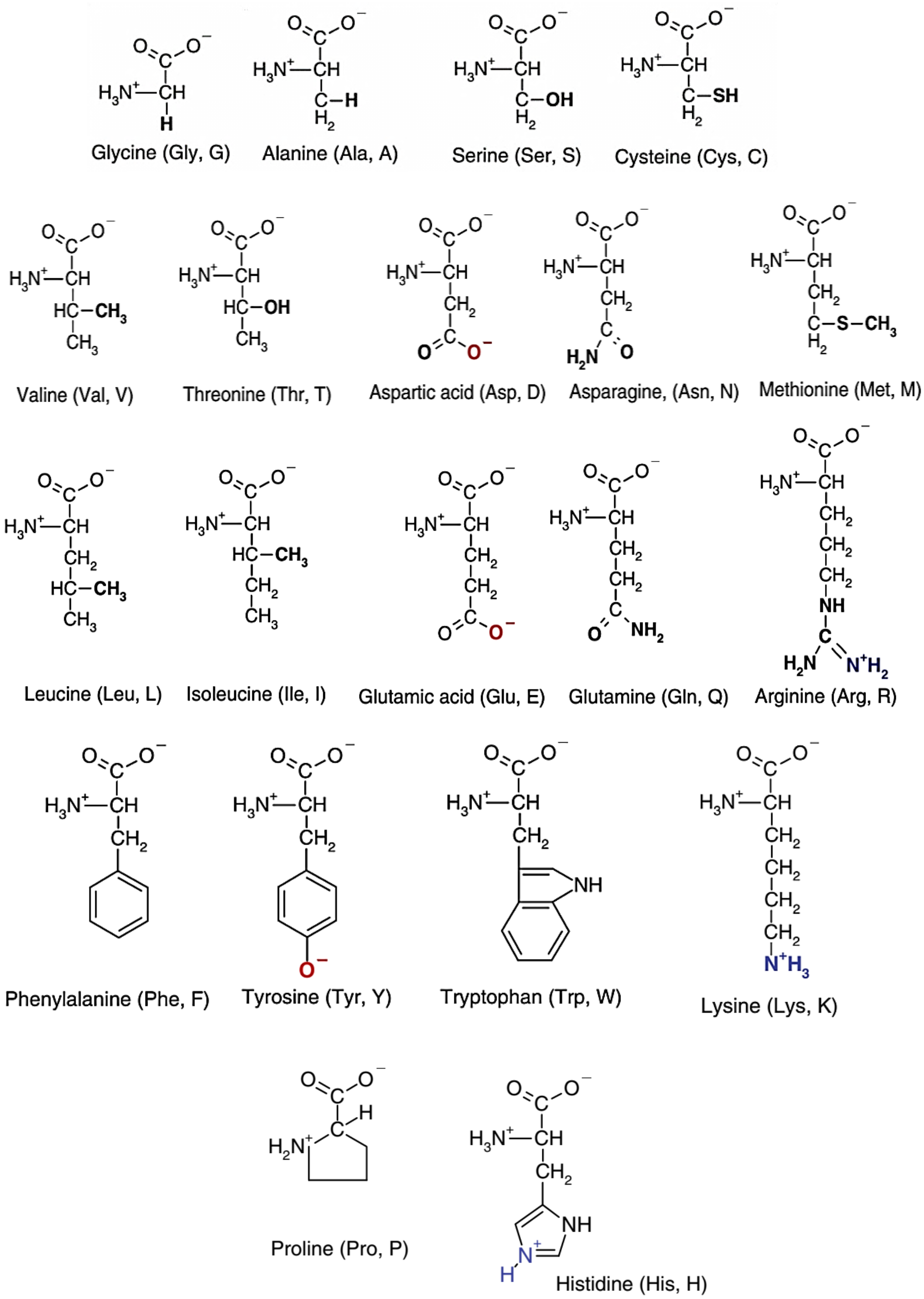

Fig. (2.2). The 20 amino acids encoded by genes, the three and one letter codes for each on. Once incorporated into proteins, amino acids may be further modified. Ile have a chiral β- in addition to the α-carbon [56].

2.4 Biological Function of Amino Acid Variety

The reason of occurring so many different amino acids is that these different molecules have different properties that let them serve different functions in proteins, Table (2.1).

There are amino acids whose side chains can bear positive or negative charges, while other side chains are always uncharged. Charged side chains have different *p*Kr, which can be influenced strongly by neighboring amino acids, for example Cys (*p*Kr = 5–10), His (*p*Kr = 4–10) and the carboxylic acid group of Glu and Asp (*p*Kr = 4–7). This is important for proton transfer reactions in the catalytic center of proteins. Ionisable groups also form the ionic bonds (salt bridges) which stabilize protein tertiary structure.

Asp, Glu and His residues can chelate bivalent metal ions like Fe, Zn and Ca. This is important for enzymes with metal co-factors, such as in hemoglobin and in some regulatory proteins like calmodulin. Some amino acids are hydrophilic (= water friendly) because they carry polar groups (-COOH, $^{-}$NH2, $^{-}$OH, $^{-}$SH). Other amino acids are hydrophobic (= water fearing, fat friendly), with long aliphatic (Ile, Leu, Val) or aromatic (Phe, Trp) side chains. If these residues point into the solution, they force water molecules into a local structure of higher order (i.e. lower entropy), which is unfavorable. Burying these residues into the interior of the protein avoids this penalty, this is the molecular basis for hydrophobic interactions.

Some amino acids have small side chains (like glycine), others very big, bulky ones (like tryptophan). The small hydrogen residue of Gly not only fits into tight spaces but because it has no β-carbon it can assume secondary structures that are forbidden for all other amino acids.

Proline has its nitrogen in a ring structure, which makes the molecule very stiff, limiting the flexibility of protein chains.

The SH-group of Cys, the unprotonated His and the OH-group of Ser and Thr are nucleophiles which are essential residues in the active centre of enzymes.

Table (2.1). Properties of the 20 amino acids encoded in a mammalian genome. Post-translational modification may change these properties considerably. The helix propensity measures the energy by which an amino acid destabilizes a poly-alanine helix [56] .

Amino acid	MW	pK1(-COOH)	pK2(NH_3^+)	pK3(Side group)	*p*I	Hydropathy	Abundance(%)
Ala	89	2.34	9.69	-	6.01	+1.8	9.0
Arg	174	2.17	9.04	12.48	10.76	-4.5	4.7
Asn	132	2.02	8.08	-	5.41	-3.5	4.4
Asp	133	1.88	9.60	3.65	2.77	-3.5	5.5
Cys	121	1.96	8.18	10.28	5.07	+2.5	2.8
Glu	147	2.19	9.67	4.25	3.22	-3.5	6.2
Gln	146	2.17	9.13	-	5.65	-3.5	3.9
Gly	75	2.34	9.60	-	5.97	-0.4	7.7
His	155	1.82	9.17	6.00	7.59	-3.2	2.1
Ile	131	2.36	9.68	-	6.02	+4.5	4.6
Leu	131	2.36	9.60	-	5.98	+3.8	7.5
Lys	146	2.18	8.95	10.53	9.74	-3.9	7.0
Met	149	2.28	9.21	-	5.74	+1.9	1.7
Phe	165	1.83	9.13	-	5.48	+2.8	3.5
Pro	115	1.99	10.96	-	6.48	-1.6	4.6
Ser	105	2.21	9.15	13.60	5.68	-0.8	7.1
Thr	119	2.11	9.62	13.60	5.87	-0.7	6.0
Trp	204	2.38	9.39	-	5.89	-0.9	1.1
Tyr	181	2.20	9.11	10.07	5.66	-1.3	3.5
Val	117	2.32	9.62	-	5.97	+4.2	6.9

Some amino acids confer properties to the protein which can be used in the laboratory: Met binds certain heavy metals which are used in X-ray structure determination and reacts with cyanogen bromide (Br−C≡N) to cleave the protein at specific sites. Cys and Lys are easily labeled with reactive probes. Aromatic amino acids, in particular Trp absorb UV-light at 280 nm, this can be used to measure protein concentration. In addition they show fluorescence, which can be used to measure conformational changes in proteins.

2.5 Protein Synthesis

Proteins are assembled from amino acids using information encoded in genes. Each protein has its own unique amino acid sequence that is specified by the nucleotide sequence of the gene encoding this protein. The genetic code is a set of three-nucleotide sets called codons and each three-nucleotide combination designates an amino acid, for example AUG (adenine-uracil-guanine) is the code for methionine. Because DNA contains four nucleotides, the total number of possible codons is 64; hence, there is some redundancy in the genetic code, with some amino acids specified by more than one codon.[57] Genes encoded in DNA are first transcribed into pre-messenger ribonucleic acid (mRNA) by proteins such as RNA polymerase. Most organisms then process the pre-mRNA (also known as a *primary transcript*) using various forms of Post-transcriptional modification to form the mature mRNA, which is then used as a template for protein synthesis by the ribosome. In prokaryotes the mRNA may either be used as soon as it is produced, or be bound by a ribosome after having moved away from the nucleoid. In contrast, eukaryotes make mRNA in the cell nucleus and then translocate it across the nuclear membrane into the cytoplasm, where protein synthesis then takes place, Figure (2.3). The rate of protein synthesis is higher in prokaryotes than eukaryotes and can reach up to 20 amino acids per second [58].

The process of synthesizing a protein from an mRNA template is known as translation. The mRNA is loaded onto the ribosome and is read three nucleotides at a time by matching each codon to its base pairing anticodon located on a transfer RNA molecule, which carries the amino acid corresponding to the codon it recognizes. The enzyme aminoacyl transfer RNA (tRNA) synthetase "charges" the tRNA molecules with the correct amino acids. The growing polypeptide is often termed the *nascent chain*. Proteins are always biosynthesized from N-terminus to C-terminus [57].

The size of a synthesized protein can be measured by the number of amino acids it contains and by its total molecular mass, which is normally reported in units of *daltons* (synonymous with atomic mass units), or the derivative unit kilodalton (kDa). Yeast proteins are on average 466 amino acids long and 53 kDa in mass [62]. The largest known proteins are the titins, a component of the muscle sarcomere, with a molecular mass of almost 3,000 kDa and a total length of almost 27,000 amino acids [59].

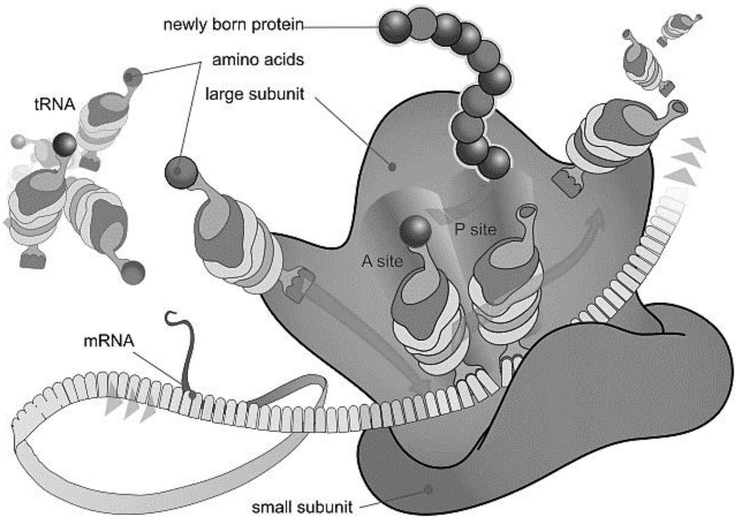

Fig (2.3) Protein synthesis process

2.6 Protein structure

The architecture of protein molecules is quite complex. Nevertheless, this complexity can be resolved by defining various levels of structural organization, Figure (2.4).

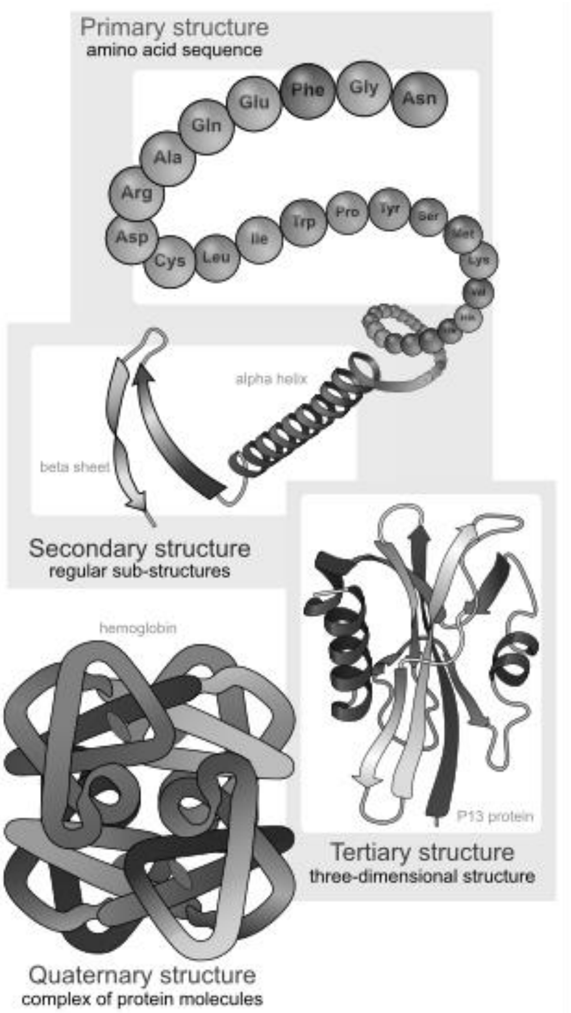

Fig (2.4) Different levels of protein structures

2.6.1 Primary Structure

The sequence of amino acids in a protein is called its primary structure. In biochemistry, this is always given starting with the N-terminal and ending with the C-terminal amino acid, because this is the order in which amino acids are added during protein synthesis in the cell [61]. Cα, C', the nitrogen and the oxygen atom of the peptide bond form a single plane. The bond between C' and N is somewhat shorter than a normal C−N single bond, because of mesomery with the C=O double bond, Figure (2.4).

No rotation is possible around double bonds, this is also true for the "partial" double bond between C' and the peptide nitrogen. Thus the R-groups of the amino acids can occur in cis - or in trans-configuration, Figure. (2.5). Because of the bulky R-groups, the trans-configuration is more stable for most amino acids (99.95 % probability). The exception is Pro, which occurs in cis - configuration much more frequently than other amino acids (6 % probability).

On the other hand, the N-Cα and Cα -C' bonds are normal single bonds, rotation around those is possible. The angles of rotation are named φ and Ψ respectively. Rotation in the peptide chain is limited by two factors. First, at certain angles φ and Ψ around one amino acid atoms of that amino acid would collide with atoms of the following amino acid. These angles are forbidden, by definition we assign the value 0° to the angle that would result in collision of C'n with Nn+1. Rotation angles then can have values between −180° and +180°. Additionally, size and charge of the R-groups can make certain positions more stable than others.

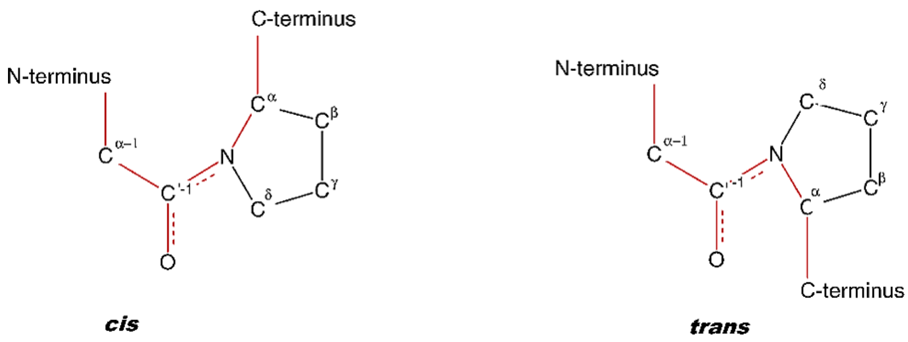

Fig (2.5). Cis-trans -isomery around the peptide bond. Because the C −1 -N-bond has the character of a partial double bond, rotation around this bond cannot occur and cis-trans isomery results. For steric reasons the Trans-configuration is much more probable

than the cis -. Pro is unusual in that the cis-configuration has a probability of 5–6 %, which is about 100 times higher than with other amino acids [56].

Thus in a plot of φ versus Ψ (Ramachandran-plot [62], [63], figure (2.5) there are regions which are sterically forbidden, there are fully allowed regions with no steric hindrance, and there are unfavorable regions which can be assumed by slight bending of bonds.

Proline is again a special case because the peptide nitrogen is part of a ring structure, this limits φ to values between −35° and −85°. As we will see in a moment, this has considerable consequences for protein secondary structure.

Because glycine has only a hydrogen as R-group, steric hindrance is much less of a problem than with other amino acids. Thus in a Ramachandran-plot Gly can be found in regions forbidden for other amino acids.

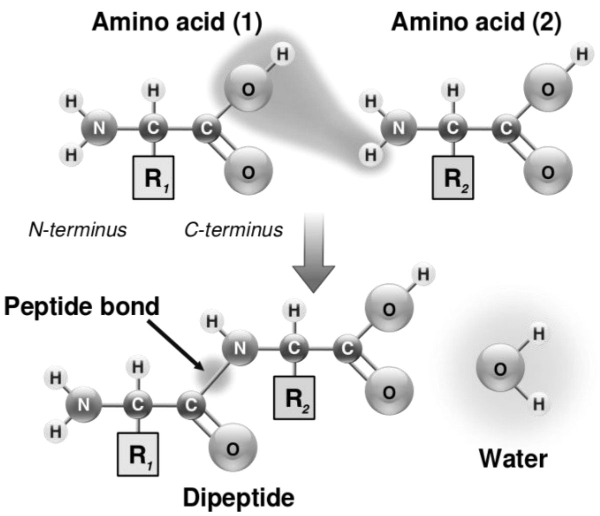

Fig. (2.6). Polycondensation of amino acids to peptides and protein. Polyconden- sation is a reaction were organic molecules react with each other via their functional groups, producing small molecules (here: water) in addition to a macromolecule [56].

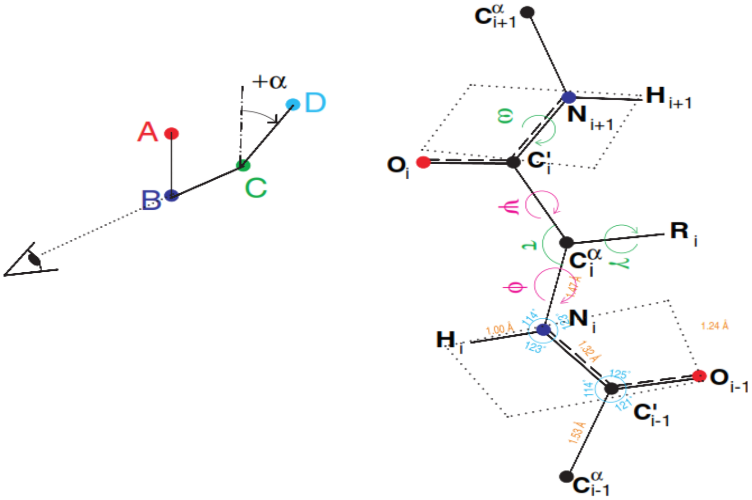

Fig. 2.7. The geometry of the peptide bond. Left: Standard way to determine the dihedral angle of a bond (here between atoms B and C). Orient that bond into the paper plane, so that the neighboring atoms (here A and D) point upwards. Then measure the angle formed, clockwise is positive, anticlockwise negative. Right: Because of mesomery, the dihedral angle of the bond between the carboxy-carbon (C') and the nitrogen (ω) is fixed to 180°, with N, H, C and O lying in a single plane. Slight deviations are possible, but rare. The bond angle of Cα (τ) is usually the tetrahedral angle 109.5°, but some flexing (± 5°) is occasionally found. Variable are the dihedral angles φ and Ψ, which determine the secondary structure of a protein [56].

2.6.2 Secondary Structure

Secondary structure describes the local conformation of the amino acids in the protein chain. It is stabilized by hydrogen bonds. figure (2.8) between the amino- and keto-groups of the peptide bonds, which carry a partial positive and negative charge, respectively, although each hydrogen bond has only a relatively small bond energy, the sum of the bond energies over all hydrogen bonds in a protein is considerable.

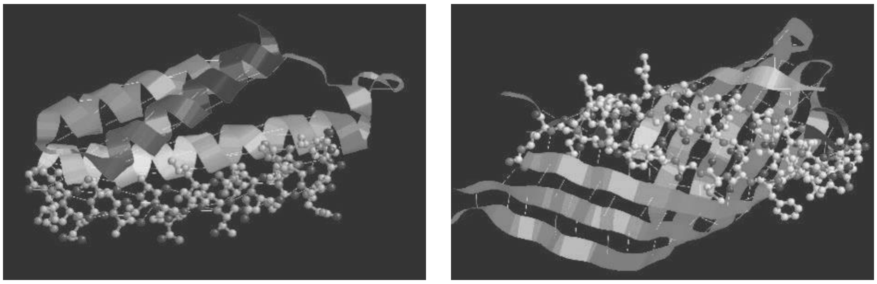

Fig. (2.8). Hydrogen bonding in α-helix (left, cytochrome b562, PDB-code 256B) and β-sheets (right, E. Coli OmpA, PDB-code 1QJP). In an α-helix all hydrogen bonds between keto- and amino-groups in the protein backbone occur between neigh- bouring amino acids of the same helix. In β-sheets however all such hydrogen bonds occur between amino acids in diﬀerent strands, alternating between the right and left neighbor [56].

There are four particularly common structural motives, which illustrated as follows.

2.6.2.1 α-helix Structure

The polypeptide chain is wound around an imaginary axis, with 3.6 amino acids per turn. Each turn is about 5.4 A° long, the pitch per amino acid Figure (2.9)

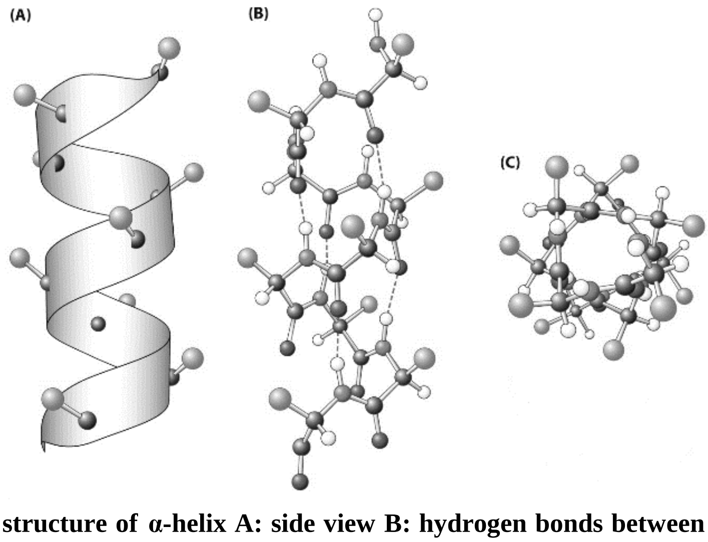

Fig. (2.9) The structure of α-helix A: side view B: hydrogen bonds between a.as C: upper view