Protein Structure & Motifs

Biochemistry 201
Molecular Biology
January 12, 2000
Doug Brutlag

Introduction

Proteins are more flexible than nucleic acids in structure because of both the larger number of types of residues and the increased flexibility and lower charge density of the polypeptide backbone. Proteins can serve many roles in the cell, as enzymes, as structural components, membrane components, as templates, as substrates and as products of reactions. Many aspects of protein metabolism are catalyzed and regulated by the cell. These include their rates of expression, their translation, their folding, their degradation and their targeting to the proper cellular location. Proteins are the end product of the genes that encode them. Some of the most important functions of proteins are to regulate the expression of other proteins.

In this lecture we will discuss the components of proteins, their covalent structure, their non-covalent interactions, higher order structures such as motifs and domains and then give several examples of different types of protein folds. It will be extremely useful for you to down load the Kinemage 4.2 program and the Proteach Kinemage collection for reviewing the material presented in the class. Pointers to the locations to obtain this program and the Proteach files are on the course Web page.

Amino Acids

The amino acid residues of proteins are defined by an amino group and a carboxyl group connected to an α-carbon to which is attached a hydrogen and a side chain group R. The smallest amino acid, glycine, has a hydrogen atom in place of a side chain. All other amino acids have distinctive R groups. Because the α-carbon of the other amino acids have four different constituents, the α-carbon atom is an asymmetric center and most naturally occurring amino acids are in the L form.

Amino acids fall into several naturally occurring groups including hydrophobic, hydrophilic, charged, basic, acidic, polar but uncharged, small polar, small hydrophobic, large hydrophobic, aromatic, β-branched, sulfur containing etc. Hydrophobic amino acids are sometimes called non-polar and reside primarily on the interior of the protein. Hydrophilic amino acids are sometimes called polar and reside primarily on the exterior of the protein. Many amino acids will fall into more than one group since each amino acid side chain has several properties.

Protein Structure

Peptide Bonds

Amino acids are linked to each other by peptide bonds. The dehydration of the carboxyl group of one amino acid and the amino group of the next form peptide bonds. Because of the resonance structure of the electron orbitals on the amino and carboxyl groups, the peptide bond is planar.

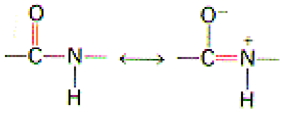
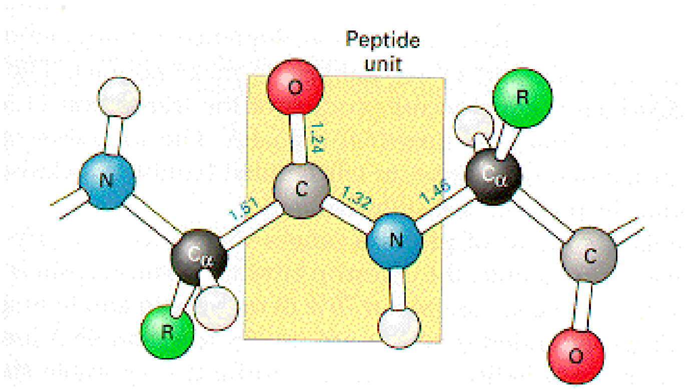
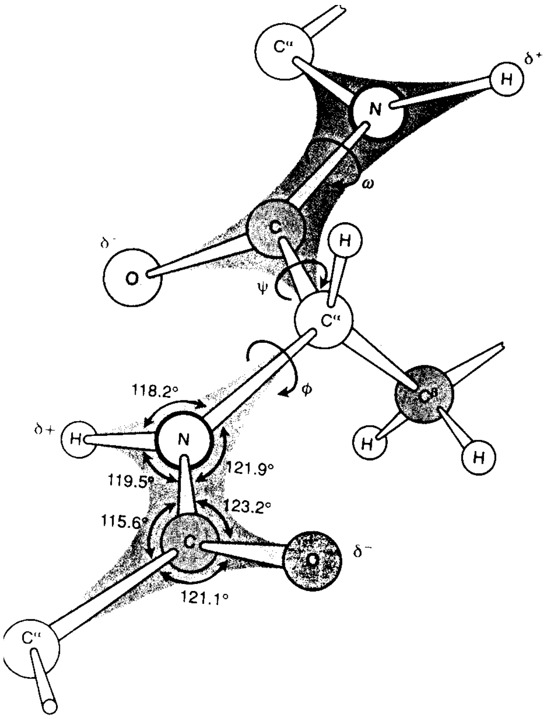


Figure 2-29
The peptide unit is planar because the carbon–nitrogen bond has partial double-bond character.





The dihedral angle between the amino group and the α-carbon and the α-carbon and the carboxyl group are free to rotate and these angles are referred to as the phi-psi angles.

Glycine, with the smallest side chain, has the most conformational flexibility about the phi-psi angles. Other amino acids are restricted in their rotation due to steric hindrance from the side chains. The rotation of the dihedral angles of side chains about the different bonds, referred to as chi-1, chi-2 etc., are also restricted for different side chain elements. Proline, which in which the side chain is linked back to the backbone is the most restricted, only two conformations are permitted.

Protein Structure

Forces determining protein structure

There are several covalent and non-covalent forces that determine protein structure. The list of forces include (not exhaustive):

1) **Van der Waals interactions between immediately adjacent atoms.** These non-covalent forces result from the attraction of one atoms nucleus for the electrons of another atom in a non-covalent form (no sharing of orbitals). These forces are much weaker than covalent interactions and the interaction distances are much longer than covalent bonds and much shorter than the other non-covalent interactions. Van der Waals interactions occur at distances between 3 and 4 Å. They are very weak beyond 5Å and electron repulsion prevents atoms from getting much closer than 3Å. Van der Waals interactions are non-directional and very weak. However, significant energy of stabilization can be obtained in the central hydrophobic core of proteins by the additive effect of many such interactions.

2) **Hydrophobic force.** The hydrophobic force is really a negative non-covalent force. The presence of hydrophobic side chains in aqueous solution induces the formation of structured water (clathrate cages of water molecule form, like miniature ice crystals about the hydrophobic side chains). This reduction in entropy of the water molecules is a very unfavorable resulting in a strong force to keep hydrophobic side chains buried in the interior of the protein. The hydrophobic force is one of the largest determinants of protein structure. Most secondary structural elements we will discuss have an amphipathic nature, one hydrophobic side and one hydrophilic side because the structure lies on the surface of the protein.

3) **Electrostatic forces.** The attraction of oppositely charged side chains can form salt-bridges, which stabilize secondary and tertiary structures. The electrostatic force is quite strong, falling off as the square of the distance between the charged atoms. It also depends heavily on the dielectric constant of the medium in which the protein is dissolved. It is strongest in a vacuum and 80 fold weaker in water and weaker still at elevated salt solutions. Water and ions can shield electrostatic interactions reducing both their strength and distance over which they operate.

4) **Dipole moments.** Dipole moments are caused by pairs of charges separated by a larger distance than permitting a salt- or ion bridge. The dipole moment can give rise to an electric field along the entire length of a structural element and are often used by proteins to attract and position charged substrates and products. The peptide chain naturally has a dipole moment because the N-terminus carries about 1/2 a positive charge and the C-terminus carries about 1/2 unit of negative charge. The α-helix is known to carry a partial negative charge at its C-terminus and a positive charge at its N-terminus. In order to help neutralize this charge distribution, α-helices often have acidic residues near their N-terminus and a basic residue near their C-terminus.

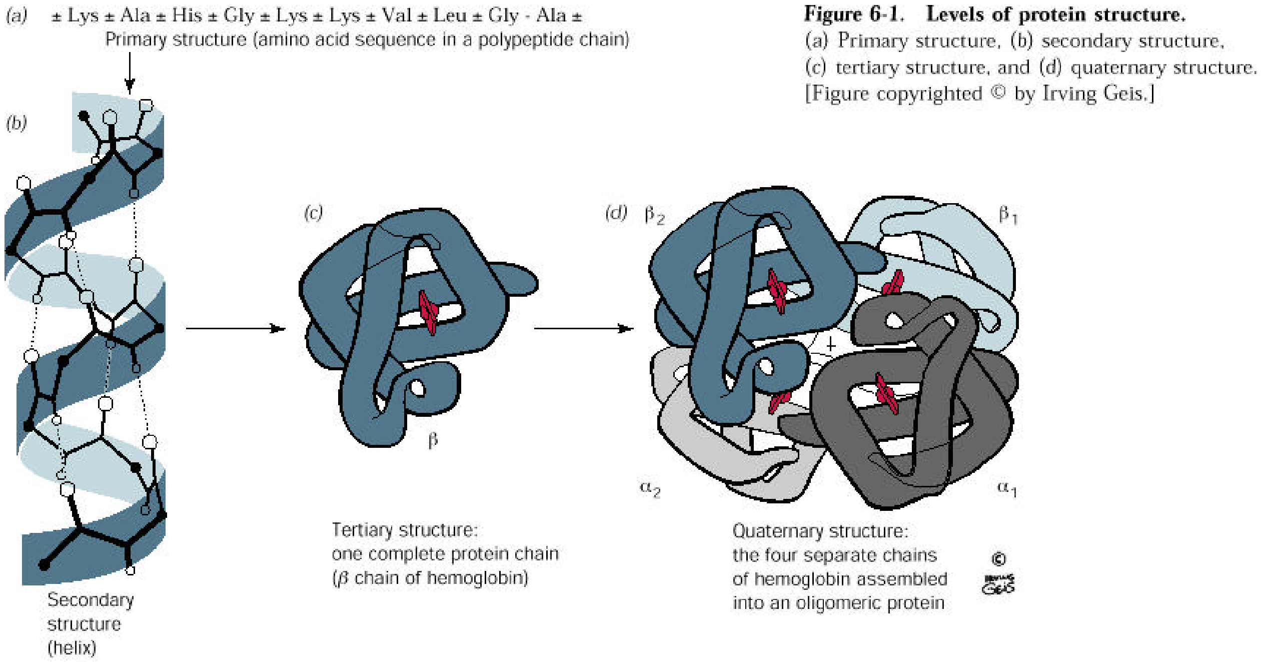
5) **Hydrogen bonds.** Hydrogen bonds occur when pair of nucleophilic atoms such as oxygen and nitrogen share a hydrogen between them. The hydrogen may be covalently attached to either nucleophilic atom (the H-bond donor) and shared with the other atom (the H-bond receptor). H-bonds are very directional and their strength deteriorates

dramatically as the angle changes. Hydrogen bonds do not, in general, contribute to the net stabilization energy of proteins because the same groups that hydrogen bond to each other in a native protein structure, can hydrogen bond to water in the denatured state. However, hydrogen bonds are extremely important because of their directionality, they can control and limit the geometry of the interactions between side-chains. This is shown most dramatically in patterns of hydrogen bonding between the carboxyl groups and the amino groups in the peptide backbone that give rise to α-helix and β-strand conformations.

6) **Covalent bond distances and torsion angles.** The major properties of the covalent bonds hold proteins together are their bond distances and bond angles. In particular, the bond angles between two adjacent bonds on either side of a single atom, or the dihedral angles between three contiguous bonds and two atoms control the geometry of the protein folding. The preferred dihedral angles for different secondary structural elements are discussed below.

Levels of Protein Structure

There are four levels of protein structure depicted below. The amino acid sequence is generally referred called the primary structure of a protein. The secondary structure is the first level of folding of the polypeptide bond and it is determined by 1) the planar nature of the peptide bond and by the phi-psi angles about the α-carbons of each amino acid. The tertiary structure of a protein is often referred to as the fold of the protein and


(a) ± Lys ± Ala ± His ± Gly ± Lys ± Lys ± Val ± Leu ± Gly - Ala ±
Primary structure (amino acid sequence in a polypeptide chain)
(b)
(c)
(d)
β
β₂
β₁
α₂
α₁
© Irving Geis

Secondary structure (helix)

Tertiary structure: one complete protein chain (β chain of hemoglobin)

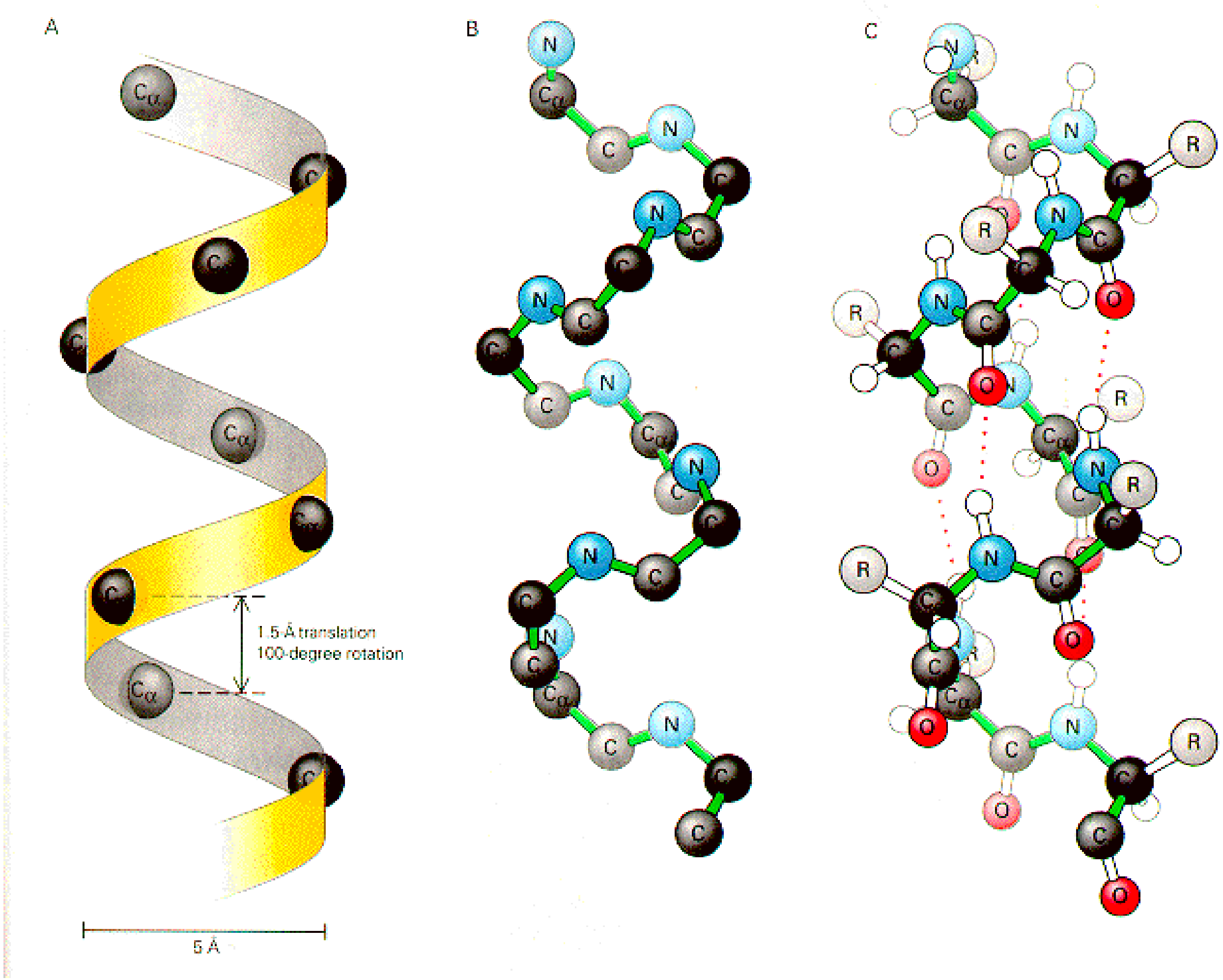
Quaternary structure: the four separate chains of hemoglobin assembled into an oligomeric protein

Figure 6-1. Levels of protein structure. (a) Primary structure, (b) secondary structure, (c) tertiary structure, and (d) quaternary structure. [Figure copyrighted © by Irving Geis.]

Protein Structure

Preferred secondary structures

α-helices are the most well known element of protein structure, proposed by Pauling and confirmed in the first structure determined, myoglobin, α-helices have distinctive patterns of hydrogen bonding and phi-psi angles. They are generally between 5 and 20 residues in length, but some proteins and coiled-coil structures can be considerably longer. The carboxyl groups of the backbone hydrogen bond to the amino group of a residue four amino acids distant along the chain. α-helices generally have a pitch of about 3.5 residues per turn, but there are forms of helices with tighter (3 residues per turn) and longer (4 residues per turn).



α-helices can be coiled about them selves in both two coil, three coil and four coil (four helix bundle) conformations. α-helices can exist internal in proteins (generally hydrophobic), on the surface of proteins (amphipathic) or in membranes (hydrophobic). α-helices can span membranes either singly or in groups.

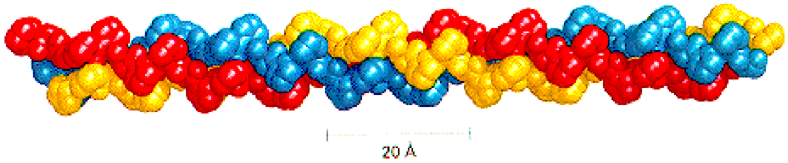
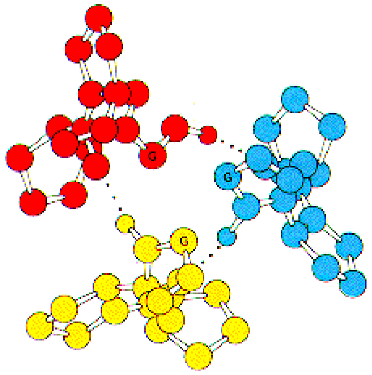
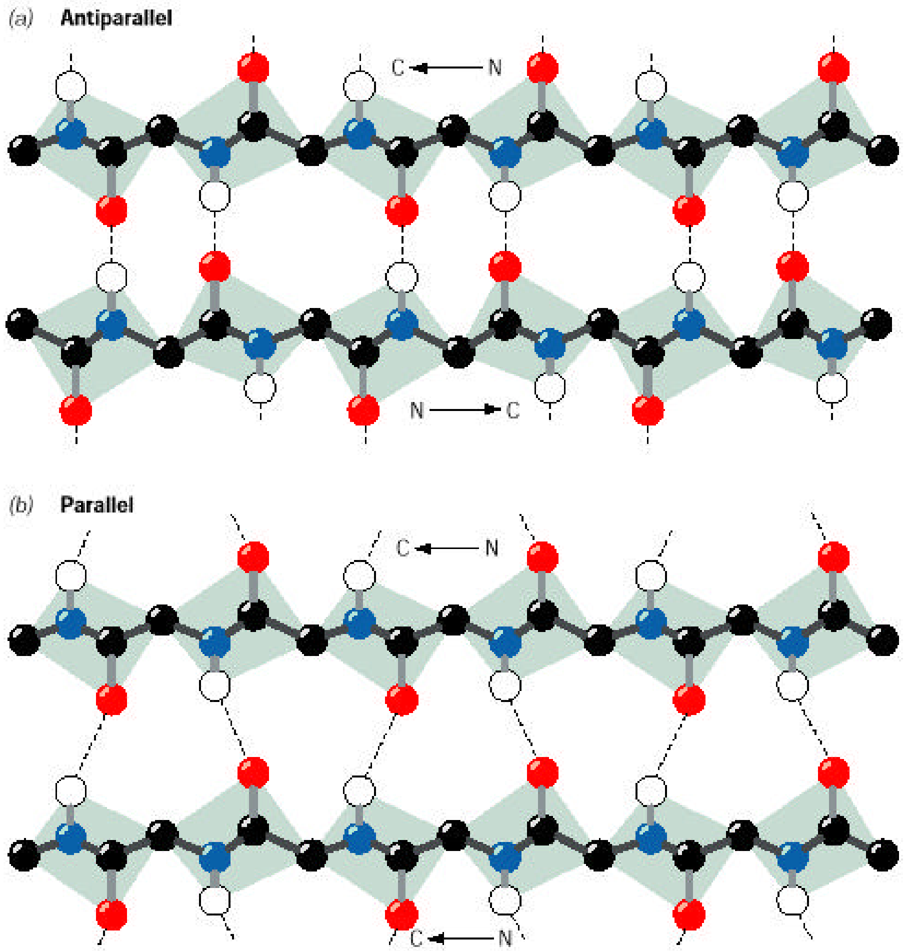


Figure 2-40
Space-filling model of the collagen triple helix. The three strands are shown in different colors. [Drawn from coordinates reported in K. Okuyama, K. Okuyama, S. Arnott, M. Takayanagi, and M. Kakudo. *J. Mol. Biol.* 152(1981):427.]

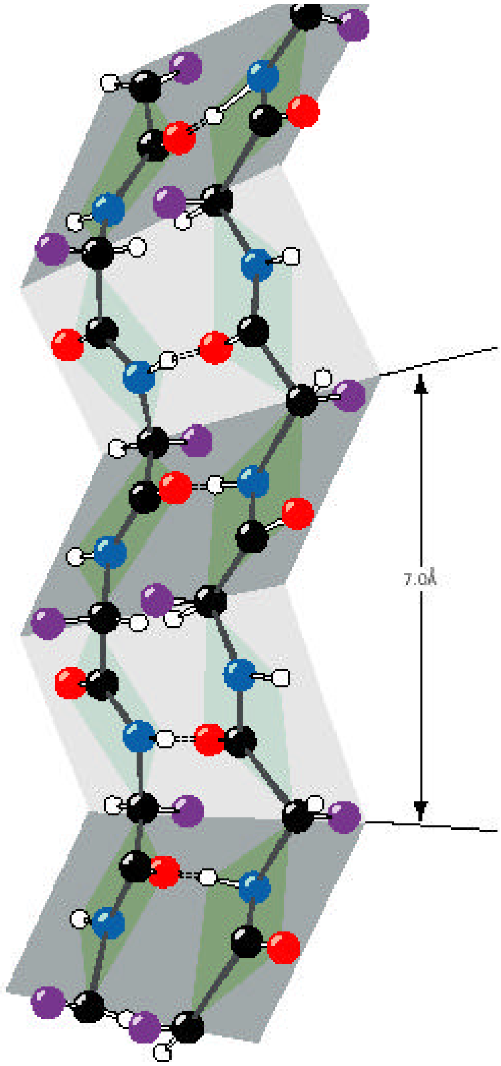


β-strands are an extended form in which the side chains alternate on either side of the extended chain. The back bones of β-strands hydrogen bond with the backbone of an adjacent β-strand to form a β-sheet structure. The strands in a β-sheet can be either parallel or anti-parallel and the hydrogen bonding pattern is different between the two forms.



Anti-parallel β-stands are often linked by short loops containing 3-5 residues in highly characteristic conformations. Longer loops are occasionally found where the loop plays an important role in substrate binding or an active site. The antigen combining site of the immunoglobulins is an important example of this.

β-sheets can be internal to a protein (largely hydrophobic) or on the surface in which case they are amphipathic, with every other amino acid side chain alternating between hydrophobic and hydrophilic nature.



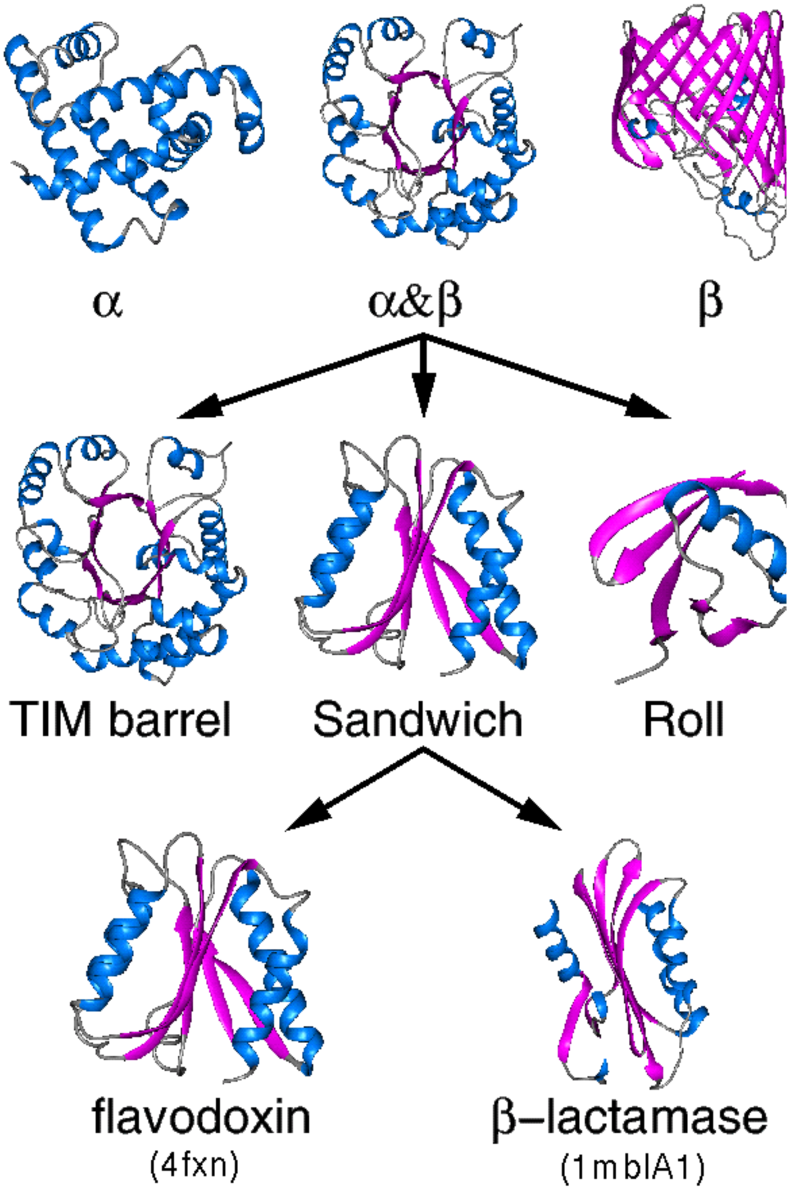
β-sheets tend to be pleated or bent as shown in the figure to the left. Every other residue points to the left or to the right side of the sheet,

Rhamachandran Plots

The peptide backbone is constrained by steric hindrance, and hydrogen bonding patterns that limit its torsional angles (phi-psi angles) to certain limits. Plots of phi versus psi dihedral angles for amino acid residues are called Ramachandran plots. One can tell if the backbone is following a helical or a an extended β-strand structure based on the values of the phi-psi angles over a length of backbone (usually 3-4 residues is sufficient).

Protein Folds or Architectures

There are a limited number of ways that secondary structures interact in the tertiary structure of proteins. These architectures or folds, as they are called are cataloged in a number of protein structure databases available on the Web. The SCOP database from MRC (also maintained here at Stanford by Steven Brenner) and the CATH database from the University College of London (Janet Thornton) are two examples. The SCOP structure database is hierarchical with four subdivisions, Class (all α-helical, all- β-strand, alternating α-helix and β-strand and α-helix plus β-strand), Fold, Superfamily and family. The later two classes often show sequence similarity but the first two levels of the hierarchy do not. Examples of the first three levels of the SCOP hierarchy are shown in the figure to the right.



All-helix or all-strand motifs

α−helices can exist independently as single elements with in proteins or as single helices spanning a membrane. More often α−helices associate with each other or with β-sheets to form a higher order of folding. Some simple examples are the coiled-coil motif such as found in the leucine zipper, triple helices as found in collagen or four helix bundles in numerous enzymes. The central core of these structures usually is composed of hydrophobic interactions of short branched aliphatic amino acids with charged pairs on the outside of the structure. The helices are wrapped around each other in a gentle right handed coil with a pitch angle of 20°.

There are other all α−protein structures such as myoglobin and hemoglobin whose helices associate in several different ways to form a heme binding site.

β-strands are almost always associated with each other to form sheets of two or more strands. The strands can be paired in a parallel or antiparallel fashion. The antiparallel strands are connected by either short turns (called β-turns) or by longer lengths called loops. Loops usually constitute active sites or binding sites for substrates of an enzyme. The antibody combining sites on immunoglobulins are composed of loops from several anti-parallel β-strands.

A single pair of antiparallel β-strands linked with a β-turn is known as a hairpin. two such pairs, if folded over in the correct conformation are known as a Greek Key. Finally if there are four or more pairs, folded so that pairs of adjacent strands intercalate between other pairs, a jellyroll motif is formed.

Several large barrel structures can also be formed exclusively from β-strands. The porin barrel which forms the structure of many metal ion permeases, are a good example of this motif.

α/β **Protein motifs.**

Several common structural motifs involve alternating β-strands and α−helices. The best known among these is the TIM barrel, named after the structure of TIM (triphosphate isomerase 1TIM in the PDB database). The TIM barrel has eight parallel strands that form a cylindrical barrel. The ends of the strands are linked together by helical segments that pass from the C-terminus of one strand to the N-terminus of the next. There are at least 17 classes of enzymes that display the TIM barrel structure. None of them appear to be related either functionally or via sequence comparisons. This is most likely an example of convergent evolution on a structure of great utility.

Another common alternating β-strand, α−helix, β-strand motif is the Rossman Fold. The Rossman fold is comprise of two copies of β-α-β-motifs. These motifs are often used to bind nucleotides in enzymes.

DNA Binding motifs.

There are many families of DNA binding motifs and only a few will be shown in class. The best understood are the zinger finger, consisting of a parallel β-hairpin and its associated α−helix, and the helix-turn-helix motif. All DNA binding proteins often have a general affinity for DNA via nonspecific, often ionic interactions with the back bone of DNA. This sometimes takes the form of an interaction between the dipole moment of a β-strand or an α−helix pointed towards the DNA binding crevice. Specificity is determined by a series of many, often greater than 10, hydrogen bonds between arginine and glutamine residues on the protein and keto groups on the bases. Many hydrogen bonds are needed to obtain the required specificity for the target sequence.

The zinc finger motif interacts with a short region of DNA, usually three or four basepairs. Most proteins that use this motif use at least three contiguous zinc fingers in order to make sufficient contacts to obtain the required specificity. It is also important that the proteins fail to bind to target sites that differ in even a single base. In order to accomplish this, often two hydrogen bonds are made to a single base pair.

Proteins that use the helix-turn-helix motif are usually dimers with two copies of the motif that bind to a region 12-15 base-pairs. They make use of the dyad axis of DNA to recognize a symmetric sequence at its target site.

Protein Sequence Motifs

Many structural motifs can often be represented as short sequence motifs. The leucine zipper for example can be represented as a "regular expression" or consensus:

L . {6} L .{6} L .{6} L .{6} L

where L refers to Leucine and .{6} means six occurrences of any amino acid. There as a large collection of such protein sequence motifs referred to as the Prosite database and it is very useful for identifying functional regions in newly sequenced proteins from the genome efforts.

Protein sequence motifs are usually generated by aligning sequences of similar function or structure and then choosing the most common amino acid in each position or a group of amino acids that would represent that position. Often this method is manual and highly biased. It also does not give any insight into the structural or functional requirements of the active sites that are being represented. A more objective and systematic approach that takes into account physical and chemical properties of the amino acids will be presented in class.

References

Bairoch, A. and Apweiler, R. (1999). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res*, **27**(1), 49-54.

Hofmann, K., Bucher, P., Falquet, L. and Bairoch, A. (1999). The PROSITE database, its status in 1999. *Nucleic Acids Res*, **27**(1), 215-219.

Brenner, S. E., Chothia, C., Hubbard, T. J. & Murzin, A. G. (1996). Understanding protein structure: using scop for fold interpretation. *Methods Enzymol*, **266**, 635-43.

Branden, C. & Tooze, J. Introduction to Protein Structure, Second Edition, Garland Publishing, New York. http://www.proteinstructure.com/

Creighton, T. E. (1993). *Proteins: Structures and Molecular Properties*. (Second Edition ed.). New York: Freeman.

Creighton, T. E. (Ed.). (1992). *Protein Folding*. New York: W. H. Freeman & Co.

Darby, N. J., & Creighton, T. E. (1993). *Protein Structure*. Oxford: IRL Press.

Protein Structure

Henikoff, S. And Henikoff, J.G. (1991) “Automated assembly of protein blocks for database searching” *Nucleic Acids Res., 19*, 6565–6572.

Holm, L. and Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res*, **25**(1), 231-4.

Hutchinson, E. G. and Thornton, J. M. (1996). PROMOTIF--a program to identify and analyze structural motifs in proteins. *Protein Sci*, **5**(2), 212-20.

Hubbard, T. J. P., Murzin, A. G., Brenner, S. E. and Chothia, C. (1997). SCOP: a structural classification of proteins database. *Nucleic Acids Res*, **25**(1), 236-9.

Nevill-Manning, C., Sethi, K., Wu, T. D. and Brutlag, D. L. (1997). Enumerating and Ranking Discrete Motifs. *ISMB-97*, **4**, 202-209.

Nevill-Manning, C. G., Wu, T. D. and Brutlag, D. L. (1998). Highly Specific Protein Sequence Motifs for Genome Analysis. *Proc. Natl. Acad. Sci. USA*, **95**(11), 5865-5871.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. and Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure*, **5**(8), 1093-108.

Schulz, G. E., & Schirmer, R. H. (1985). *Principles of Protein Structure*. New York: Springer-Verlag.

Stryer, L. (1995). *Biochemistry*. (Fourth Edition ed.). New York: W. H. Freeman & Co.