Chapter 1: Protein Sequence and Structure

A living biological cell is a teeming factory of molecules of various types. Protein molecules are the factory workers and machines that produce the factory's output. Some proteins also form the factory's internal structures. DNA molecules, which encode information, serve as the blueprints for how to make the products of the factory. Fatty acid and carbohydrate molecules are the energy sources that run the factory. Lipid molecules make the factory walls. In a human cell, there are approximately 30,000 different types of proteins. In simpler biological systems, even fewer functions are required to create a living organism.

Molecular function usually arises from molecular structure. Different proteins have different functions because proteins have different native structures. *Native structures* are the 3-dimensional compact shapes that proteins adopt under physiological conditions. A protein is a linear polymer molecule that has a very large number of *conformations.* The set of all such conformations is called the *conformational space*. The native structure of the protein is only one of these many structures. The process by which the protein adopts the single native structure from among the large conformational space is called *folding* (see Figure I.3.1).

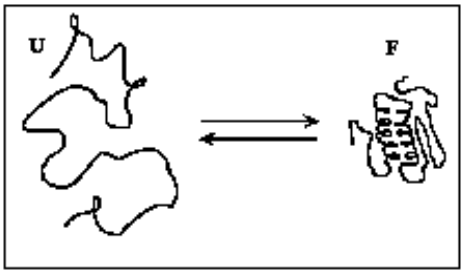


Figure I.3.1. *The protein folding process. A protein chain can adopt a large number of conformations, called the "denatured" or "unfolded" state, which then "collapses" into a single "native" or "folded" state under biological conditions.*

An analogy for protein folding is the way string, which is 1-dimensional, can ball up to form a 3-dimensional compact structure. However, the ball-of-string analogy fails in an important respect. A protein is "informational"; a string is not. Imagine a pearl necklace. Suppose you have 20 different colors of pearls. If you string the pearls together in different color sequences, you can encode information in the same way that stringing together letters from a 26-letter alphabet encodes meaning in the English language. Proteins are linear chains of monomer units. The monomers are a "20-letter alphabet" called amino acids.

A striking difference between proteins and small molecules – and probably essential to any form of life – is that structure, function, and information are all encoded within a single type of molecule. For example, benzene does not have the ability to encode 30,000 different structures and functions. The reason that proteins can encode so many different functions within a single type of molecule is because each protein's function is dictated by its 3-dimensional structure, which is *encoded* within its 1-dimensional monomer sequence.

A typical protein might contain tens to thousands of amino acids, they are covalently linked together in a specific order in one protein, say lysozyme. This sequence is different than in another type of protein, say ribonuclease. A protein that has fewer than about 20 amino acids is typically called a *peptide.* The sizes of proteins relative the sizes of cells and other biological structures are shown in figure I.4.6

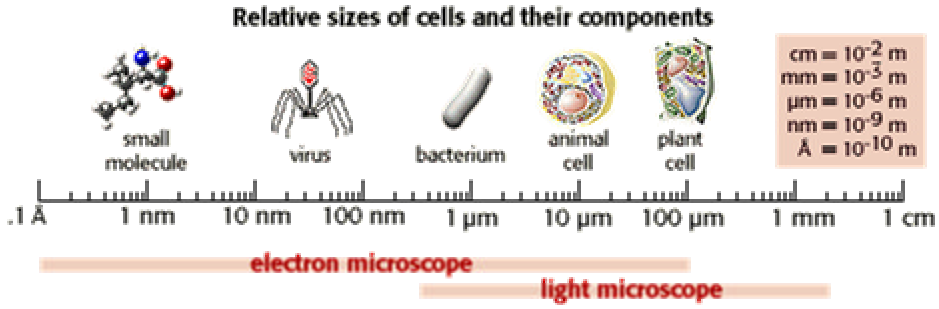


Figure I.4.6. Relative sizes of molecules, proteins, viruses, bacteria, animal and plant cells. (**http://www.biology.arizona.edu/cell_bio/tutorials/cells/cells2.html**)

(I) Proteins Are Strings of Amino Acids

Proteins are polymer molecules that can adopt many different internal conformations. The terms *conformation* or *configuration* will be used interchangeably in this book. In organic chemistry, the term conformation has a different meaning than the term configuration. However, in polymer theory and statistical mechanics, conformation and configuration are often used interchangeably (Flory, 1969). In this text, we follow the latter convention. We use the term conformation to refer to forms that a macromolecule can adopt without covalent bond breakage. We follow the standard statistical mechanical usage for the term configuration, which long predates its usage in organic chemistry, by which we mean any distinguishable state of a system accessible on the appropriate time scale. Thus, the configuration of a polymer system is identical to its conformation; and for small molecules, the configuration refers to the translational, rotational, vibrational and other accessible states.

(a) The monomer units are amino acids

The monomer units, or building blocks, of proteins are the *amino acids.* An amino acid is so-named because it contains an *amino* group ($-NH_2$) at one end, and a carboxylic *acid* (–COOH) at the other end. At physiological pH both the amino and carboxylic groups are completely ionized. The amino acid can thus act as either an acid or a base. The amino and carboxylic groups are joined by a carbon atom, called the C^{α}-atom, or α-carbon, hence the name α-amino acid. Combined, this unit is called the *backbone* of the amino acid. Each amino acid contributes two *polar* groups, -N-H and –C=O, to the protein backbone, when it is in the linear chain. Attached to the C^{α}-atom is a hydrogen atom and an amino acid *sidechain*, often denoted by the symbol "R", when referring generically to any sidechain.

Twenty natural amino acids make up the overwhelming majority of proteins. See *Figure I.1.2* and *Table I.1.* It is also possible to have other amino acids – compounds with the same backbone, but with different sidechains. Such *non-natural* amino acids are occasionally found in proteins. An example is shown in Figure I.x.x.

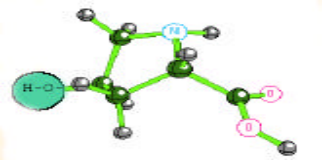


Figure I.x.x. Hydroxyproline. *Post-translationally, some prolines are modified at the C4 position by prolyl hydroxylase, by adding a hydroxyl group. This modification can enhance solubility and stability. Because ascorbic acid (Vitamin C) is utilized in this synthesis, a deficiency causes scurvy in which prolines are not modified, and collagen denatures more readily, even at normal body temperature.* *http://www.biochem.emory.edu/classes/BAHS501/2003/LectureNotes/LectureNotes16.htm*

Table I.1.1 Amino acids naturally occurring in proteins

Amino acid Type	Three-letter code	Single letter abbreviation	Occurrence[(a)] (%)
Alanine	Ala	A	8.3
Alanine	Arg	R	5.7
Asparagine	Asp	N	4.4
Aspartic acid	Asp	D	5.3
Cysteine	Cys	C	1.7
Glutamine	Gln	Q	4.0
Glutamic acid	Glu	E	6.2
Glycine	Gly	G	7.2
Histidine	His	H	2.2
Isoleucine	Ile	I	5.2
Leucine	Leu	L	9.0
Lysine	Lys	K	5.7
Methionine	Met	M	2.4
Phenyl alanine	Phe	F	3.9
Proline	Pro	P	5.1
Serine	Ser	S	6.9
Threonine	Thr	T	5.8
Tryptophan	Trp	W	1.3
Tyrosine	Tyr	Y	3.2
Valine	Val	V	6.6

[(a)] Frequency in 1021 unrelated proteins of known sequence {McCaldon & Argos 1988 ID: 289}, also plotted as the ordinate in this chapter's Appendix Figure I.1.1.

Amino acids have *chirality*, or *handedness*: they are not superimposable on their mirror image, just as a left hand is not superimposable on a right hand. This is a typical property of molecules containing a *chiral* carbon atom, – a

tetrahedral carbon having four different substituents, such as the C^{α}-atoms of amino acids (except glycine). *Figure II.2.1* shows the two possible *optical isomers* called the *L-isomer* – also known as the left-handed (*levo*) form, or the *D-isomer* – the right-handed (*dextro*) form of amino acids. These rotate the plane of plane-polarized light in opposite directions. For some reason, biology evolved using almost exclusively L-amino acids. Note that these two isomers are fundamentally of different character from the bond rotational isomers of macromolecular chains, in that there is no possibility of interconversion between the D- and L- isomers, except by breaking chemical bonds.

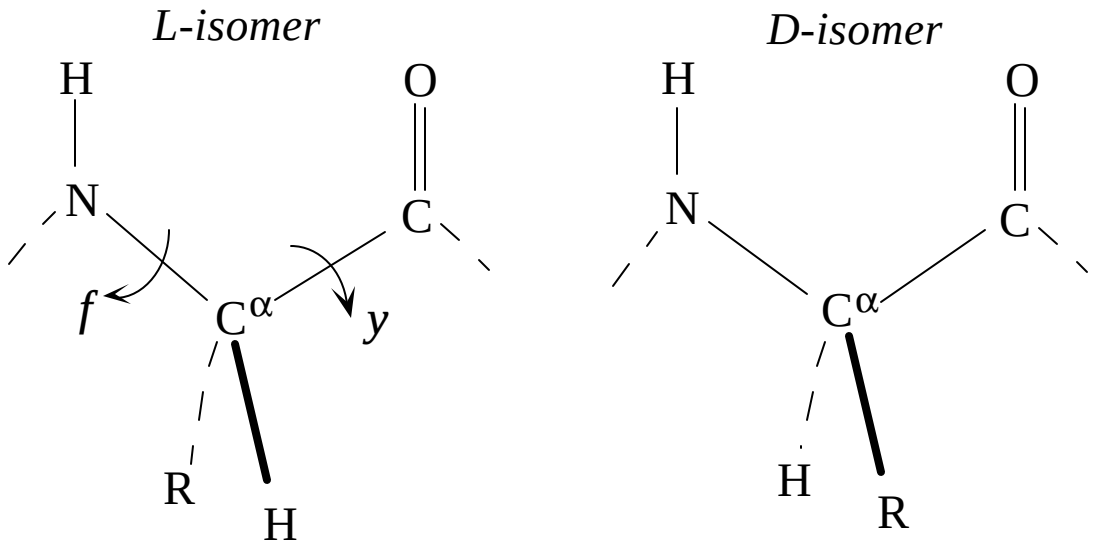


Figure II.2.1. *Two optical isomers, L- and D-, for α-carbons of amino acids. Note the difference in the relative position of the sidechains R. Dashed and boldface lines designates the bonds pointing behind and in front of the plane spanned by the backbone bonds N-C^{α} and C^{α}-C. The side groups N-H and C=O lie in the plane of the backbone bonds. The L-isomer is the form naturally occurring in proteins. The two backbone bonds flanking the α-carbon are rotatable. Their rotational angles f and y are indicated (on the L-isomer).*

b. Different Amino Acids Have Different Physical Properties

Table I.1.1 shows the volumes of amino acids, and their frequencies of occurrence in proteins. Depending on the chemical properties of their sidechains, amino acids are grouped into classes, such as *charged*, *polar* (P), and *nonpolar*. Lysine and arginine are positively charged; aspartic acid and glutamic acid are negatively charged. Alanine, valine, leucine, isoleucine, and phenylalanine sidechains are *nonpolar*, since they are comprised of hydrocarbons only, $-CH_2-$ (methylene) and $-CH_3$ (methyl) groups, which are not readily polarized by electrical fields. Glycine may also be put into this group, having only a hydrogen atom, the smallest possible 'sidechain', affixed to its C^{α}–atom. Nonpolar groups have an aversion to water. Nonpolar groups tend to cluster together and form contacts among themselves to avoid water. Amino acids having nonpolar sidechains are *hydrophobic* (H), except for glycine and alanine, the two smallest amino acids in which the backbone polar groups dominate the behavior. On the other hand, cysteine and methionine, despite their polar sulfur atoms, tend to be buried in the interior of proteins, similarly to hydrophobic residues. Serine and threonine are *polar* because of their –OH (hydroxyl) terminal groups; and asparagine and glutamine are polar because of their $(C=O)-NH_2$ groups at the sidechain termini. These have an affinity for water.

These groupings are neither precisely defined nor fully accurate. Several amino acids have multiple personalities, having both nonpolar and polar or ionic character. Most of these groups have ionization states determined by their intrinsic pKa values together with the local pH. For example, tyrosine has a bulky hydrophobic part, its phenyl moiety, and a polar group, –OH at the sidechain terminus. Likewise, tryptophan, with its indole group containing two rings, is closer in character to the group of hydrophobic amino acids than the polar ones. The histidine sidechain, a heterocyclic imidazole, can ionize at physiological pH, and thus could be assigned either to the group of charged amino acids or to the group of uncharged ones, depending on the local pH. Lysine is classified as a charged residue because its terminal amino group is ionized under most physiological conditions, but its sidechain also contains a hydrophobic segment of four methylene groups. Likewise, the arginine sidechain contains three methylene groups.

Size, shape and flexibility are important properties, almost as important as hydrophobicity and polarity, in determining the behavior of amino acids. The

charged sidechains Lys and Arg having bulky and highly flexible sidechains, for example, are fundamentally different from the charged residue Asp. Lys and Arg are almost exclusively solvent-exposed, wobbling in the surrounding solution thus increasing the solubility of the protein and usually being targets for enzyme/DNA recognition; whereas Asp can accommodate inner positions, the entropy cost of restricting its conformational freedom being much lower than that of Lys or Arg. Phe, a hydrophobic sidechain, has much in common with the other aromatic sidechains Tyr and Trp. Even the residues having aliphatic sidechains, differ in their flexibility, the position of the branching of the sidechain being an important determinant of flexibility. Val and Ile are branched at their β–carbon, which confers some stiffening in the main chain; whereas no main chain hindrance occur with Leu because of its branching at the more distant γ–carbon. Gly is unique. Its small size aids in accommodating tightly packed regions, while maintaining an intrinsic flexibility due to the rotational freedom of the corresponding (ϕ,ψ) angles.

The Venn diagram in *Figure II.2.2* summarizes some of these chemical and physical properties. It is clear that amino acid substitutions depend closely on the extent of the physical and chemical properties that they share, and this Venn diagram can thus be correlated with amino acid conservation/mutation expectancies.

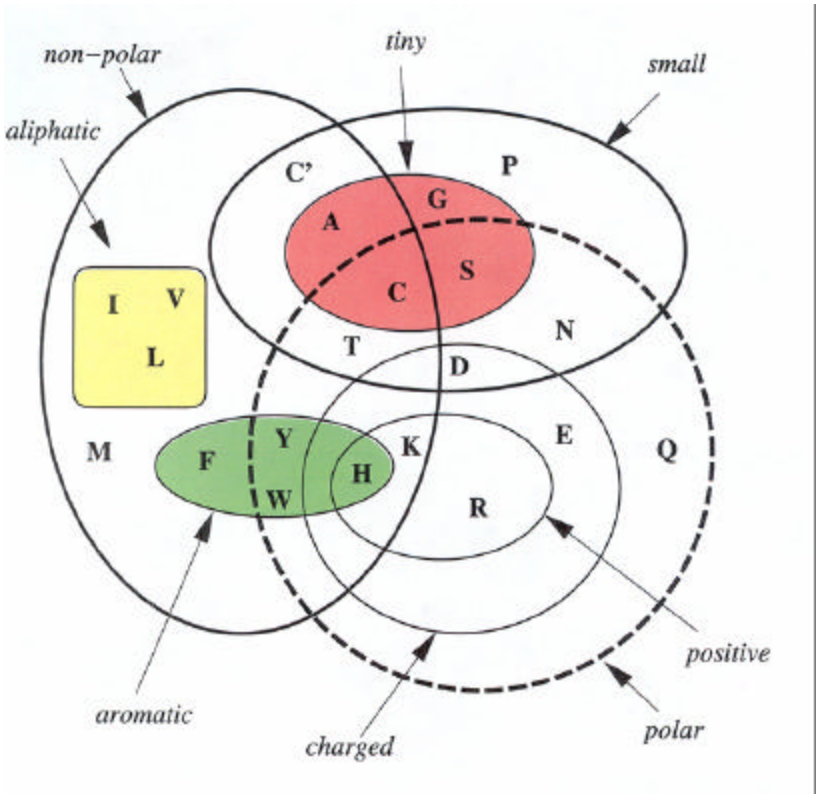


Figure II.2.2. *Venn diagram illustrating the physical and chemical similarities and differences of amino acids A similar diagram was proposed by Taylor (Taylor, 1986).*

c. In Proteins, Amino Acids are Connected by Peptide Bonds

In proteins, an amino acid is covalently connected to its neighbor by a *peptide bond*, hence proteins are sometimes called *polypeptides*. The peptide bond is a covalent link from the carbonyl end of one repeat unit –[NH –C$^{\alpha}$HR–CO]–, to the amino at the beginning of the next. The chemical process of forming a peptide bond is a *condensation* reaction, because it involves the release of one water molecule. The formation of a polypeptide of n repeat units (or *residues*), on the other hand, releases n-1 water molecules. The corresponding reaction is

$$n(HNH\text{–}C^{\alpha}HR\text{–}COOH) \longrightarrow H\text{–}[NH\text{–}C^{\alpha}HR\text{–}CO]_n OH + (n\text{-}1)H_2O$$

In conformity with the above head-to-tail connection of repeat units, the protein ends are referred to as the N- and C-terminal ends, and amino acids along the chain are assigned serial indices 1, 2, …., n starting from the N-terminus.

Figure II.2.3 displays a segment of two amino acids i and i+1 in the *trans* conformation of the backbone. For simplicity, only the C^{β} atoms are displayed. Three types of backbone bonds are distinguished, contributed by each residue: N – C^{α}, C^{α} – C and the peptide bond C – N. The peptide bond is quite rigid, with its directly attached carbonyl oxygen, amide hydrogen and two successive C^{α} atoms essentially on the same plane (*Figure II.2.4*). Its rigidity originates in its having partial double bond nature. The two backbone bonds flanking the C^{α} carbons, on the other hand, are free to rotate about their own axis. The backbone may thus be viewed as a succession of planar groups (peptide units) that are relatively free to pivot around the C^{α} carbons. This distinctive structural feature permits us to express the backbone configuration in terms of *virtual bonds* that join successive α–carbons. See *Figure II.2.5.*

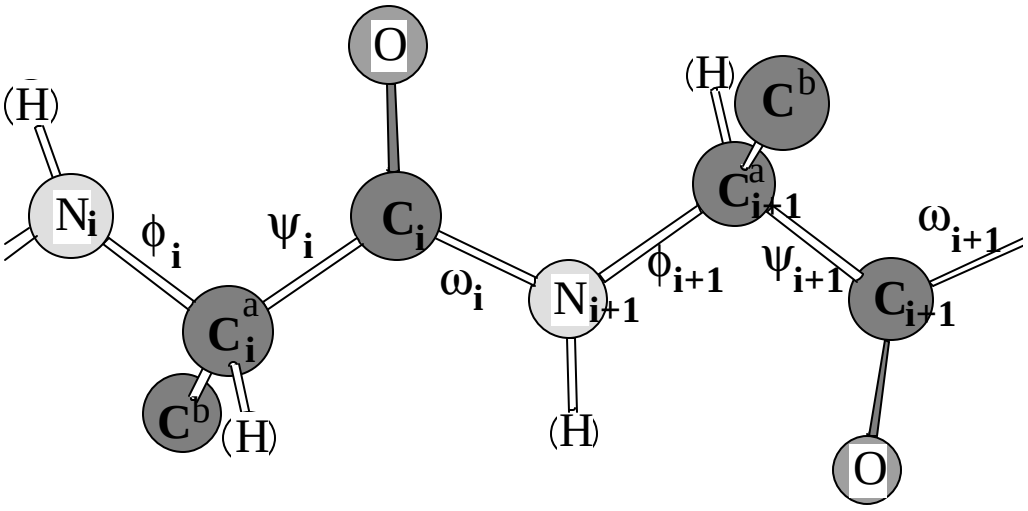


Figure II.2.3. *Segment of a polypeptide chain. Residues i and i+1 are displayed. Backbone atoms are indexed by the residue they belong to. For clarity only the C^{b} atoms of the sidechains are shown. Backbone torsions associated with the bonds $N_i - C_i^{a}$, $C_i^{a} - C_i$ and $C_i - N_{i+1}$ are denoted as f_i, y_i and w_i, respectively. The ith peptide bond connects the atoms C_i and N_{i+1}. Usually w is trans (w = 180^0) but occasionally it can be in the cis form (w = 0).*

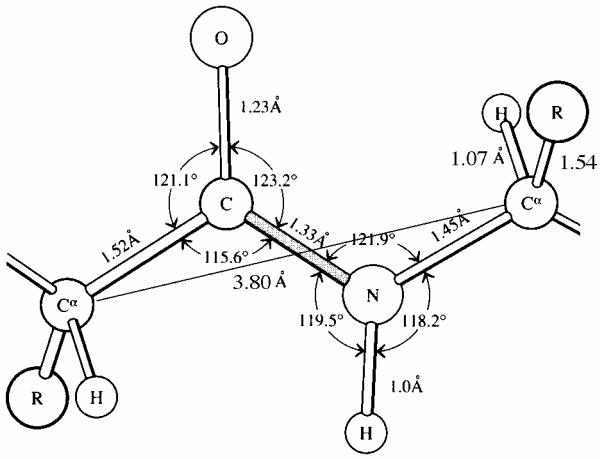


Figure II.2.4. *Geometry of the protein backbone with a peptide bond in the trans conformation, showing two adjacent C^a atoms (adapted from Figure 1.2 in (Creighton, 1993). Original ref: G.N.Ramachandran et al. Biochim. Biophys. Acta 359, 298-302, 1974)).*

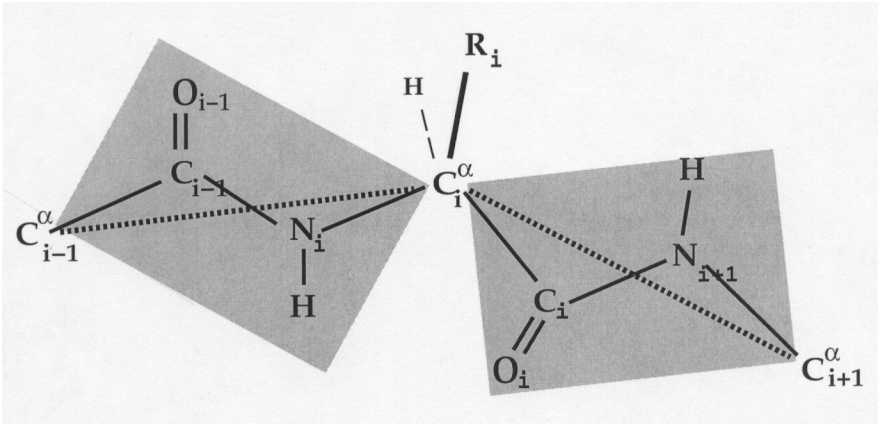


Figure II.2.5. *Virtual bond model representation of the protein backbone. Dotted lines are the virtual bonds connecting successive a-carbons. This representation takes advantage of the planarity of the three successive backbone bonds, C^{a}_{i-1}-C_i, C_i-N_i and N_i-C^{a}_{i} corresponding to each amino acid. The lengths of the virtual bonds are fixed to the extent that the bond lengths and bond angles are fixed and the amide bond adheres to the trans conformation.*

The rotational freedom around the backbone is described by two dihedral angles, f and y. The rotation around the N–C$^{\alpha}$ bond defines the torsional angle f, and that around C$^{\alpha}$– C defines y. The peptide bond rotational angle is designated as ω. It usually assumes the *trans* planar form of the peptide unit, - except for the peptide bond preceding a proline which is more likely to adopt the *cis* state because of the constraints imposed by its neighboring proline sidechain, a five-membered pyrrolidine ring.

d. Cysteines can form covalent disulfide bonds

e. Ramachandran Maps Describe the Accessible Ranges of *f* and *y* Angles

Some values of *f* and *y* are more favorable than others. Also, different amino acids have different *f y* angle preferences. Ramachandran and coworkers were the first to show that the main determinant of the different preferences of different amino acids is the degree to which the particular amino acid's sidechain interferes with the backbone by steric collision (Ramachandran et al., 1967; Ramakrishnan and Ramachandran, 1965; Ramachandran et al., 1966; Ramachandran and Sasisekharan, 1968). In their pioneering work, the permissible values of *f* and *y* were determined for Ala and Gly, using fixed bond lengths and bond angles (*Figure II.2.4*), and atoms interacting via *hard-sphere potentials* (*Table II.2.1*). The resulting so-called *Ramachandran maps* are shown *Figure II.2.6*. These are two-dimensional plots of accessible pairs of *f* and *y* angles for the ith residue, considering atoms from C^{α}_{i-1} to C^{α}_{i+1}, inclusive, the peptide bonds being constrained to their planar *trans* conformations.

TABLE II.2.1. Interatomic Distances for Hard-sphere Potentials(*)

Atom Pair	Minimum Allowable Separation(Å)	Outer Limit(Å)
CC	3.2	3.0
CO	2.8	2.7
CN	2.9	2.8
CH	2.4	2.2
OO	2.8	2.7
ON	2.7	2.6
OH	2.4	2.2
NN	2.7	2.6
NH	2.4	2.2
HH	2.0	1.9

(*) used in Ramachandran plots (Ramachandran et al., 1967). See *Figure II.2.6.*

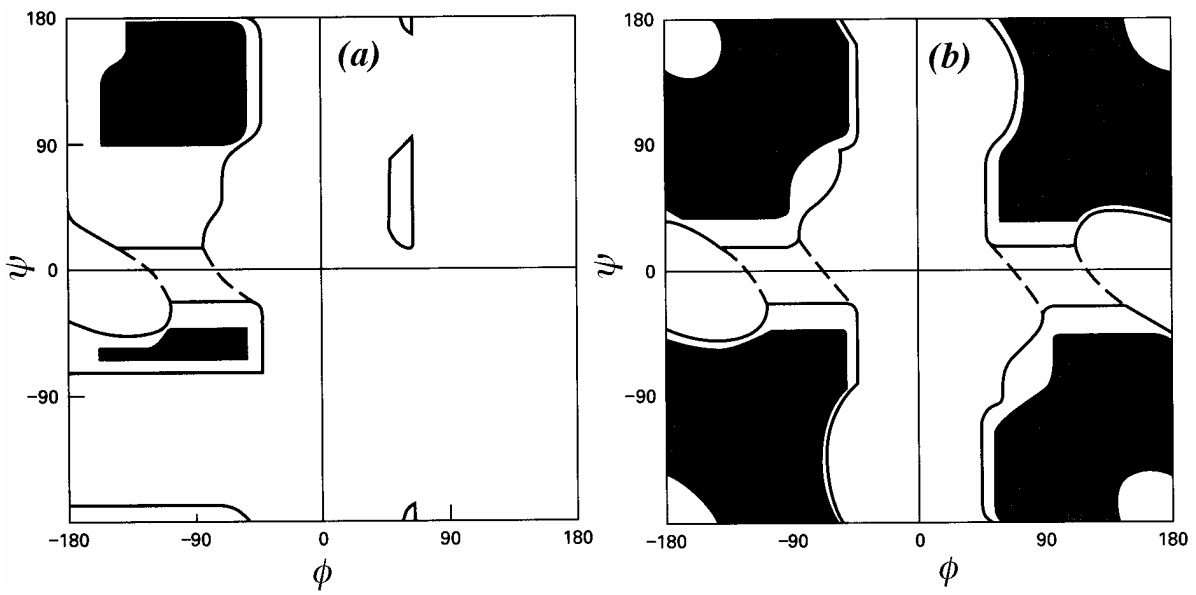


Figure II.2.6. *Ramachandran plots of the adjacent dihedral angles f and y in (a) alanine, and (b) glycine. Black regions represent the allowed regions on the basis of the fixed bond lengths and bond angles, and hard-sphere potentials with distance parameters listed in the second column of Table II.2.1. Regions enclosed by the solid lines comply with the outer limiting separations listed in the third column of Table II.2.1. The dashed curves refer to boundaries that may be reached by slight distortions in bond angles and bond lengths. The presence of chiral C^a atoms in Ala (and in all other amino acids) is responsible for the asymmetric distribution of dihedral angles in part (a), and the presence of C^b excludes the portions that are accessible in Gly. (adapted from Page 175 of (Creighton, 1993); original ref (Ramachandran and Sasisekharan, 1968)).*

Figure II.2.6 shows three favorable regions of torsional angles: one is called the *alpha* region (region I) which includes the α-helical conformations (see below), another is the *beta* region (region III), which includes parallel and anti-parallel β-sheets and the collagen triple helix conformation, and the third is the left-handed α-helix region (region II). The presence of the methyl side group discriminates between regions I and II because of the different sizes of the groups N-H and C=O. f_i values in the range $f_i < 0°$ bring the $(CH_3)_i$ and $(N–H)_{i+1}$ groups into proximity without overlap of their van der Waals radii,

while $f_i > 0°$ give rise to an overlap between $(CH_3)_i$ and the larger polar group $(C{=}O)_{i-1}$. The lower right quadrant is almost entirely precluded due to the steric clash between $(CH_3)_i$, $(C{=}O)_{i-1}$ and $(N{-}H)_{i+1}$.

The preference of *L*-polypeptides for right-handed α-helices (region I) over left-handed ones (region II) can be rationalized with regard to the larger size of the carbonyl O compared with the amide H, that restricts the access to region II. However, in proteins, the preference for right-handed α-helices is even stronger than that indicated on these Ramachandran maps. As will be shown in § II.3.a, α-helices essentially owe their stability to the hydrogen bonds formed between the polar groups $(C = O)_i$ and $(N\text{-}H)_{i+4}$, an interaction that is not included in the Ramachandran maps. Likewise, region III is favored by the hydrogen bonds formed between strands in β-sheets. Thus, Ramachandran maps are useful for providing information on the *intrinsic* preferences of amino acids in polypeptides, or on the ranges of *accessible or allowed* dihedral angles; however, the precise *distribution* of dihedral angles in proteins can, and does, depend on non-local interactions (see *Figure I.4.6*).

The above theoretical results are confirmed by experimental data. *Figure II.2.8* shows the (f, y) populations observed in proteins averaged over all amino acids except glycine and proline. *Figure II.2.9* shows the populations for some individual amino acids. Glycine and proline are unusual. Because the sidechain of glycine is only a hydrogen atom, glycine is unusually flexible, in conformity with the maps shown here. Glycine populates the left-handed helical region much more than other amino acids. Proline, on the other hand, is unusually rigid because the backbone is locked down by a small hydrocarbon ring that constitutes the side chain. Relative to other amino acids, *Figure II.2.9* shows that proline populates a very tightly defined set of f and y angles.

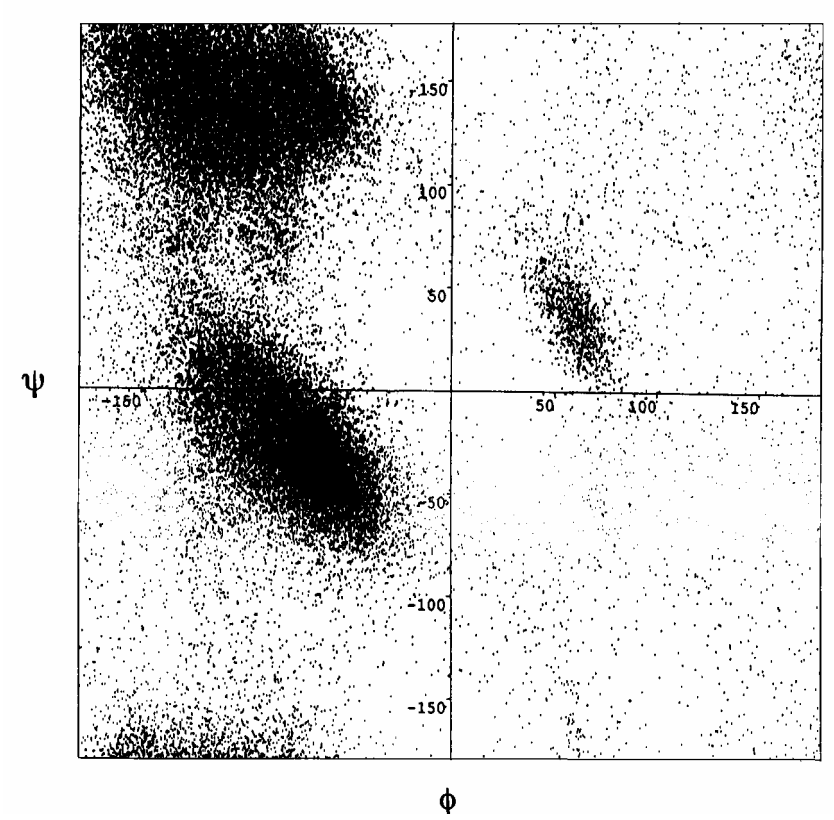


Figure II.2.8. *Ramachandran map of known protein native structures showing the (f, y ?distribution of all residues, except glycine and proline. Dots represent the observed (f, y) pairs in 310 protein structures in the Brookhaven Protein Databank (adapted from (Thornton, 1992))*

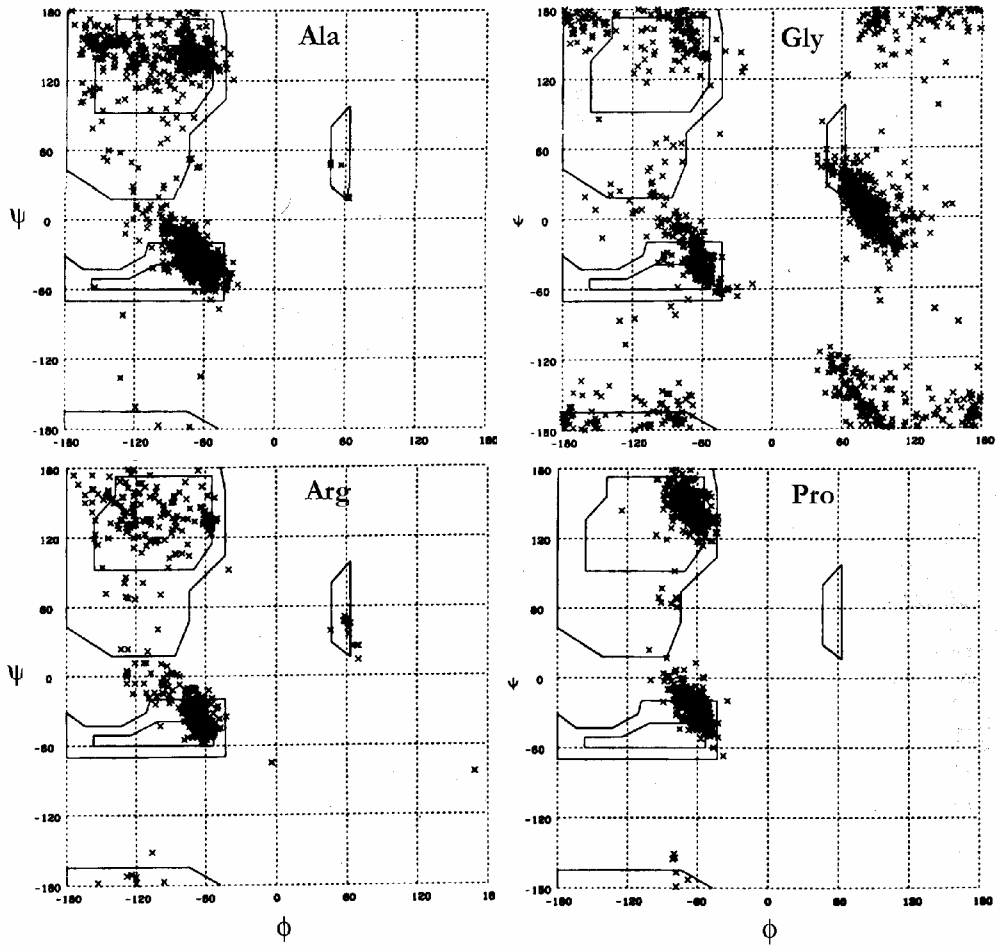


Figure II.2.9. *Ramachandran map of main chain conformations for four residue types, showing the distribution of (f, y ?angles for alanine (top left), arginine (bottom left), glycine (top right) and proline (bottom right) in high resolution protein X-ray structures. The allowed areas based on hard sphere atom calculations on alanine with two different sets of assumed hard-core radii are shown by the contours (see Figure II.2.6). See the highly expanded area around glycine and its essentially symmetric distribution about f = 0 and y = 0. The distribution for Pro is the most highly confined, with f restricted to be near -60° because of the ring (adapted from (Richards, 1992))*

f. The Virtual-Bond Model Describes the Backbone Conformations

In low-resolution approaches, it is important to preserve as much as possible of the unique, diverse characteristics of different residues, while removing the atomic details. The *virtual bond approximation* yields an almost unequivocal description of the backbone conformation, and a reasonable account of residue specificities. The distribution of bond angles and torsional angles in the virtual bond model are highly correlated with the local secondary structure, and thus reflect the secondary structure propensities of individual amino acids. See § VII.x.

A total of 2n variables define the backbone geometry for the virtual bond model: the dihedral angle (φ_i) of each virtual bond, and the angle (ϑ_i) between successive dihedral angles. The virtual bond lengths are almost fixed at 3.81 Å, except for *cis* proline. The angles ϑ_i and φ_i follow interdependent probability distributions specific to each type of residue. The coupling of the bond angles and torsion angles originates in the torsions ϕ and ψ about the backbone. Likewise, consecutive virtual bonds' dihedral angles are strongly coupled to each other and closely correlated with backbone secondary structure. *Figure II.2.10* displays the joint distribution of φ_i and φ_{i+1}. Two regions are highly populated, located near (φ_i, φ_{i+1}) ~ (50°, 50°) and ~ (200°, 200°), characteristic of α-helices, and β-sheet structures, respectively. The dihedral angles associated with β-sheets exhibit a broader distribution compared with those corresponding to α-helices, in conformity with the (ϕ, ψ) maps.

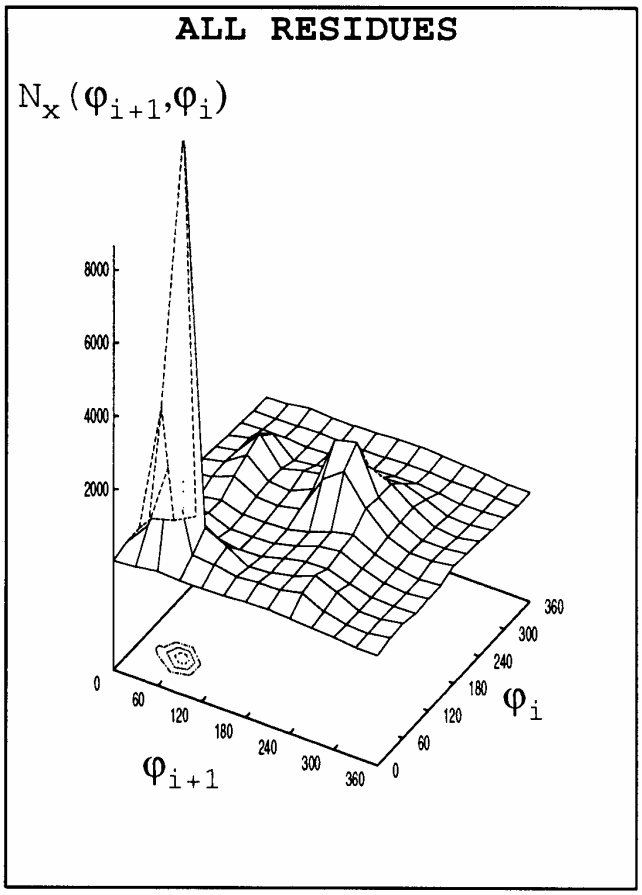


Figure II.2.10. *Distribution of adjacent virtual bonds' dihedral angles $N\chi(j_i, j_{i+1})$. The horizontal axes are the torsion angles j_i and j_{i+1}, and the surface represents the number of occurrence of each region of size ($j_i \pm 15°$, $j_{i+1} \pm 15°$) in 150 PDB structures. The projection of the surface is shown on the lower plane. The most populated dihedral angle pairs are enclosed by the innermost contours. The sharp peak near (j_i, j_{i+1}) = (50°, 50°) corresponds to a-helices, and the broader peak near (j_i, (200°, 200°) is associated with b-sheets. (from (Bahar et al., 1997))*

T he bond angle ϑ_i is found to be strongly correlated with the torsions φ_i and φ_{i+1} of the adjoining bonds. Bond angles of about 90° and 120° are favored in α-helices and β-sheets, respectively.

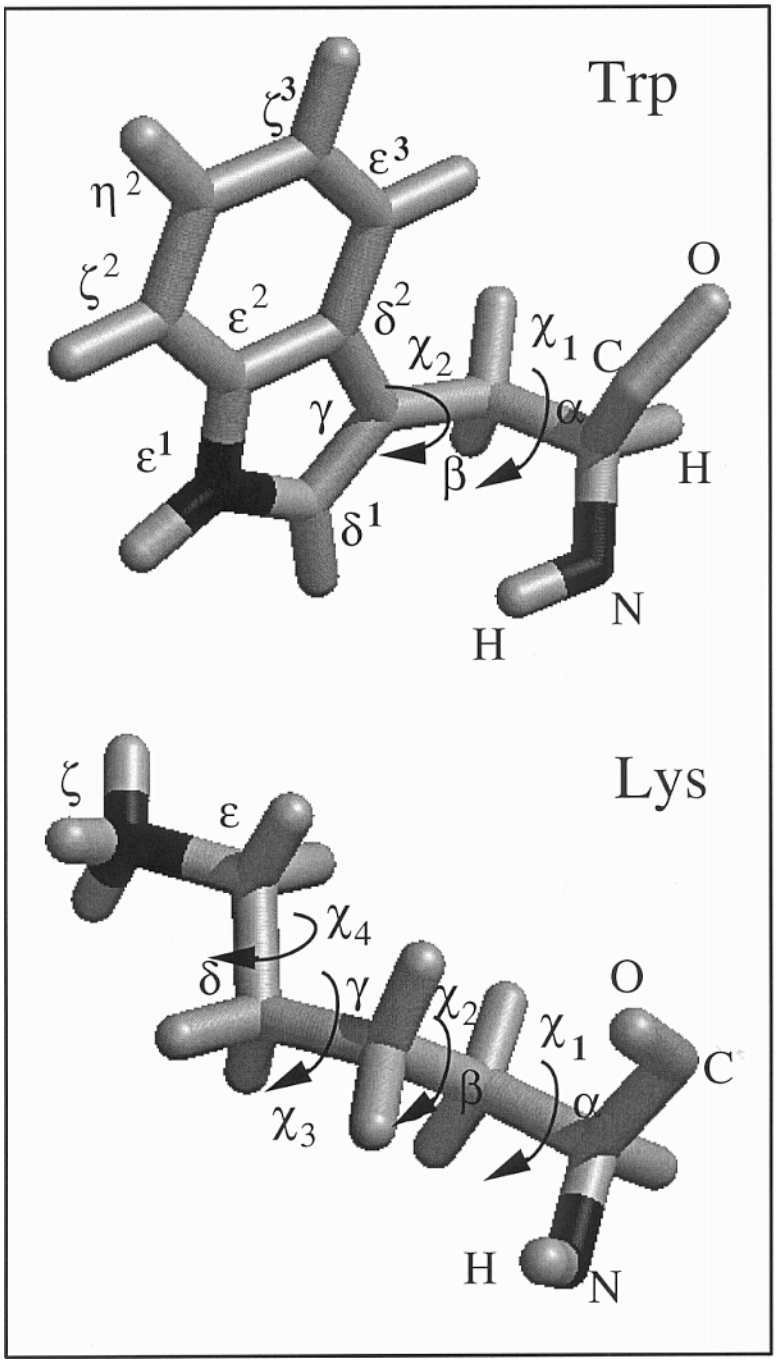
g. Sidechains Usually Adopt Preferred Conformations

In addition to f and y angles, proteins also have freedom in the side chain rotational angles. Moving along the side chain away from the backbone defines carbons identified as Cβ, Cγ, etc., and rotational angles as χ_1, χ_2, etc.; see *Figure II.2.15.*

Hydrocarbon chains *[(-CH_2-)$_n$]* in the gas phase tend to populate the three rotational isomeric states *trans*, *gauche*$^+$ and *gauche*$^-$ (see *Figures II.1.4* and *5*). *Figure II.2.16* shows that the side chains in proteins also tend to populate the same rotational angles, at least for the χ_1 angles. Figure *II.2.17* shows that this correspondence between side chain angles observed in proteins with their ideal values grows stronger as the structural quality of examined proteins increases. This is usually interpreted to mean that when proteins are known to very high resolution, side chain angles in globular proteins should coincide closely with the angles intrinsically favored by those bond types. The less optimistic interpretation is that protein side chain angles become ideal because computerized structural refinement methods are based on assuming such ideality, and that protein structure refinements reflect the artifacts of the refinement process.

But if it is true, observations of f y, and χ angles in proteins indicate an important principle that holds at least to first approximation: the folding forces acting on proteins do not perturb the backbone and side chain angles very much relative to the intrinsic values that amino acids and dipeptides would have in the absence of the folding forces. Local factors - a side chain interacting with its own backbone, or a methylene group interacting with a neighboring methylene in a side chain - dictate the bond angle options that are available to a protein. A folding protein, chooses from among those options. Folding forces rarely distort or strain the angles the backbone and side chain bonds intrinsically prefer, but simply act to select among the intrinsically favorable forms.

(A)


Trp
ζ3
ε3
η2
ζ2
ε2
δ2
χ2
χ1
C
O
γ
α
ε1
β
H
δ1
N
H
Lys
ζ
ε
χ4
δ
γ
χ2
O
χ1
C
β
α
χ3
H
N

(B)

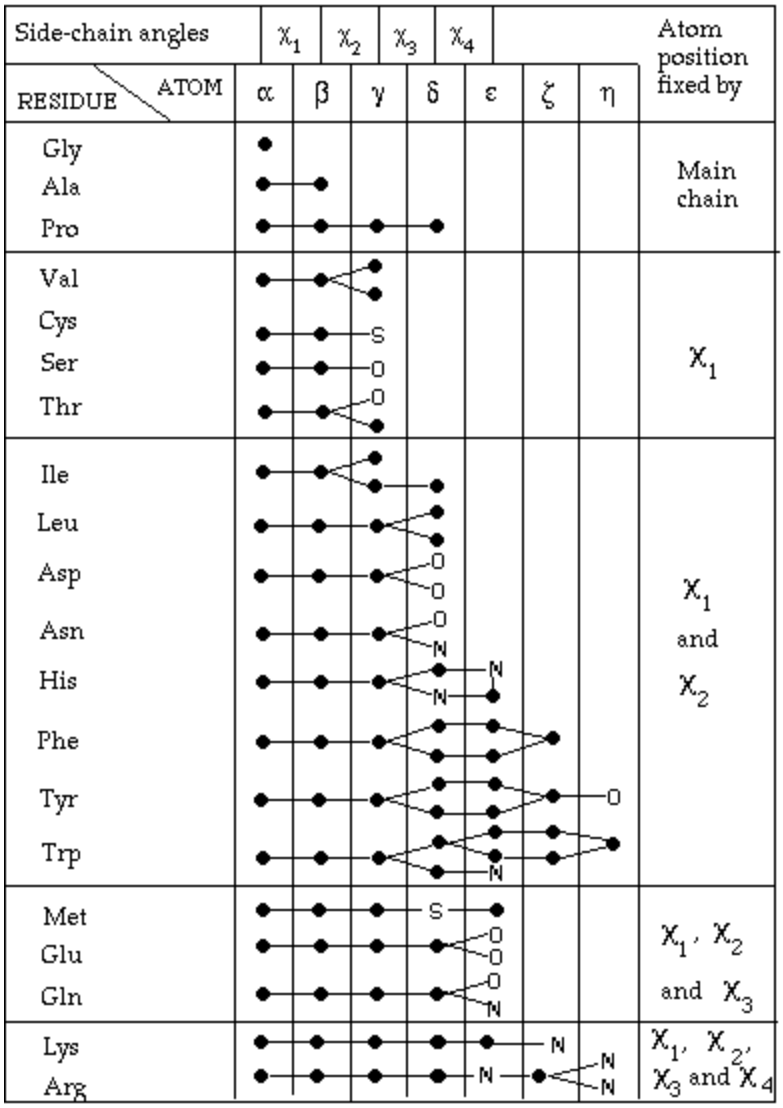


Figure II.2.15. *(A) Two examples illustrating the definition of sidechain torsional angles* c_1,c_2, *etc. Nitrogen atoms are shown in black. The labels distinguish sidechain atoms. (B) Flexibility of amino acid sidechains, showing c angles required to fix the positions of all sidechain atoms in the given residue. See ref (Ponder and Richards, 1987).*

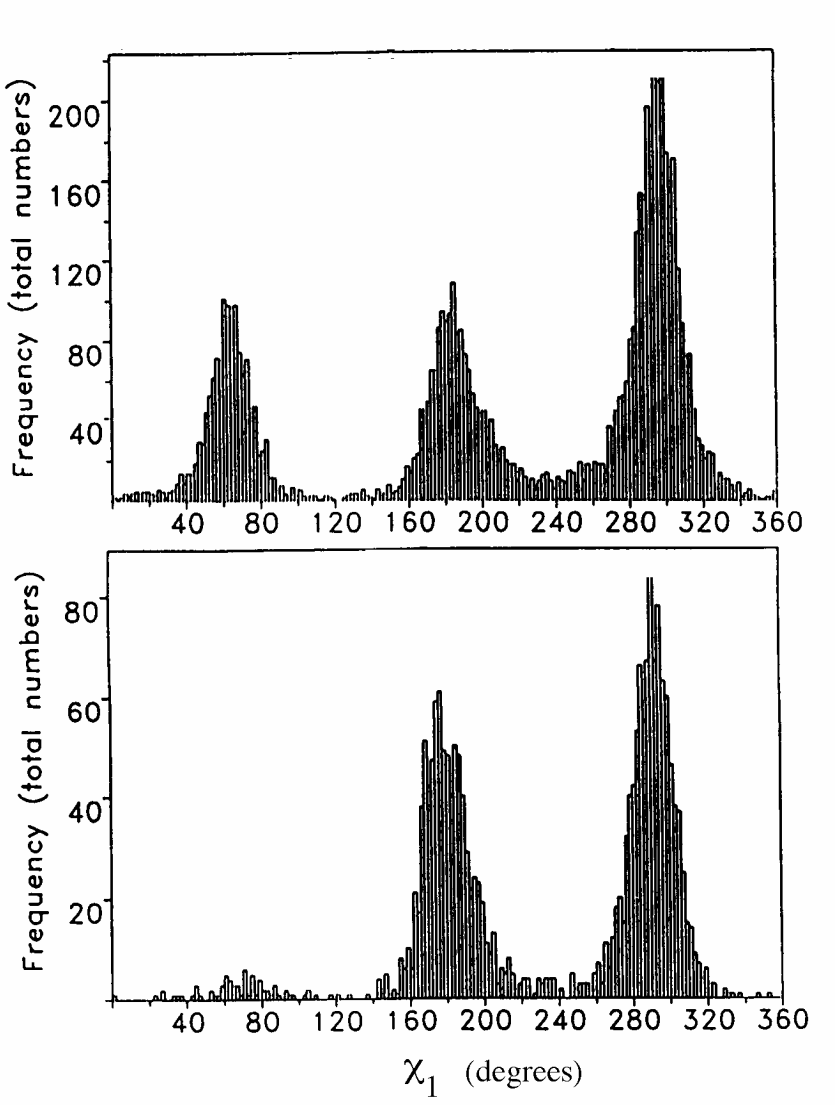


Figure II.2.16. *(A) Comparison of the c_1 sidechain torsion angle for helix and "coil" regions, compiled from 61 high resolution protein structures. (A) Distribution of c_1 values in "coil" regions (not in a-helices or b-strands), using 6,521 residues. Half of the residues belonging to the gauche$^-$ region ($c_1 = 60^o$) are serine and threonine. (B) Distribution of c_1 values among 1,972 residues in the interior of helices. These residues are so defined not to include the three residues nearest either terminus of a helix. (adapted from (Thornton, 1992))*

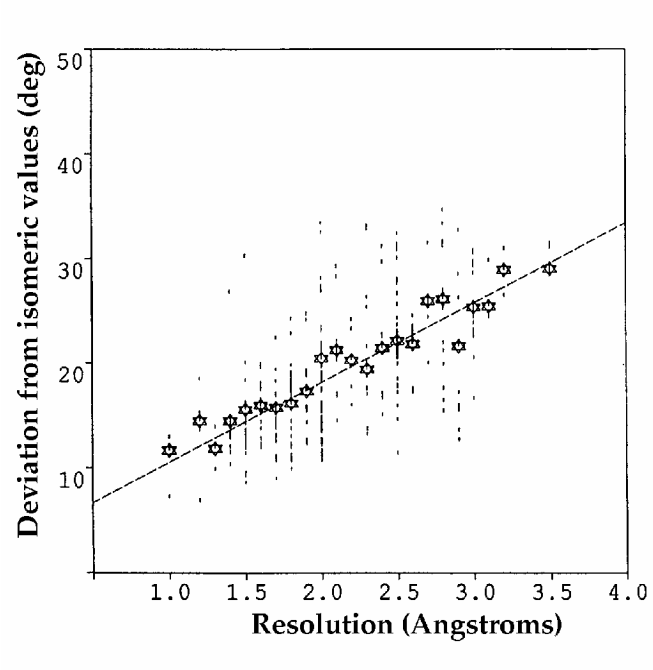


Figure II.2.17. *Deviations of c_1 angles from the standard gauche$^-$, trans and gauche$^+$ rotamers plotted against the resolution of the experimentally determined protein native structure. Points are for individual proteins, stars are averages for a given resolution. (adapted from (Thornton, 1992))*

(II) Proteins Have Hierarchies of Structure

Protein structure is regarded as having five different levels (*Fig. II.3.1*). The first level, called the *primary structure,* is the linear sequence of the amino acids in the chain. Proteins differing in primary structure differ in their amino acid sequences. The *secondary structure* describes two common patterns of structural repetition in proteins: the coiling up into *helices* of segments of the chain, and the pairing together of *strands* of the chain into β-*sheets.* The *supersecondary structure* refers to commonly observed structural patterns composed of two or more secondary structural elements folded into well-defined motifs. The *tertiary structure is* the next higher level of organization, the overall arrangement of secondary structural elements. The *quaternary structure* describes how different polypeptide chains are assembled into complexes.

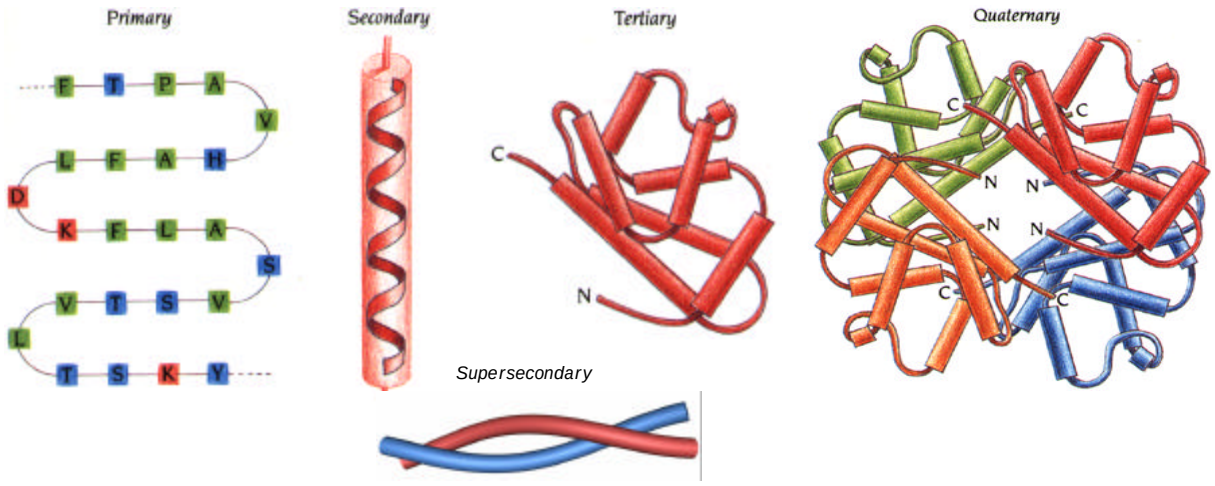


Figure II.3.1. *Different levels of protein structure. A protein chain's primary structure is its amino acid sequence. Secondary structural elements are the helices, sheets, and turns. Supersecondary structures are frequently observed motifs composed of two or more secondary structural elements. Here a superhelix formed by two helices wound into a two-stranded coiled coil is shown as an example. Tertiary structure is a polypeptide chain's three-dimensional native conformation, or the organization of the secondary and supersecondary structural elements into a unique compact structure. The example given in the figure is a schematic drawing of one of the four polypeptide chains (subunits) of hemoglobin, the protein that transports oxygen in the blood. a-helices are represented as cylinders. N is the amino terminus and C is the carboxyl terminus of the polypeptide chain. Quaternary structure is the arrangement of multiple polypeptide chains (subunits) to form a functional biomolecular structure. The figure shows the quaternary arrangement of four subunits to form the functional hemoglobin molecule. (adapted from (Branden and Tooze, 1999))*

a. Secondary Structures: Helices and Sheets are Common Motifs

a-Helices. One natural conformation for a polymer molecule is a helix. If the elementary units of a chain have the same fixed angle between every sequential pair of monomers, it defines a repeating pattern; simple geometry dictates that it will be a helix (which is three-dimensional), or a planar zig-zag (which is two-dimensional), or, if the angle is zero, a straight line (which is one-dimensional). Many *homopolymers* have a single favored repeat angle (at low temperatures) and so they crystallize into helical or zig-zag structures. Of the 176 crystal structures of polymers known in 1979 (Tadokoro, 1979), 79 were helices of 22 different types and 49 form in-plane

zigzags. The helical pitch is dictated by the physical structure of the monomer unit.

Polypeptides also form helices. The most common type of helix in proteins is called the α-helix. Discovered by Linus Pauling and his colleagues in the early l950s, the α-helix was unexpected at the time because proteins were then believed to have a very high degree of symmetry, yet the number of monomer units per helical turn predicted in the Pauling α-helix is not an integer. The defining feature of the α-helix is the backbone hydrogen bonds formed between the carbonyl oxygen of amino acid i and the amide hydrogen of amino acid i + 4 (see *Figure II.3.2*). Therefore, the α-helix has 3.6 amino acids per turn. α-helices can be formed by any of the amino acids because the hydrogen bonds are among backbone atoms, not side chains, and therefore helices are only weakly affected by side chain type, with the exception of proline that is missing a proton on its backbone N atom. Most helices in proteins are relatively short.

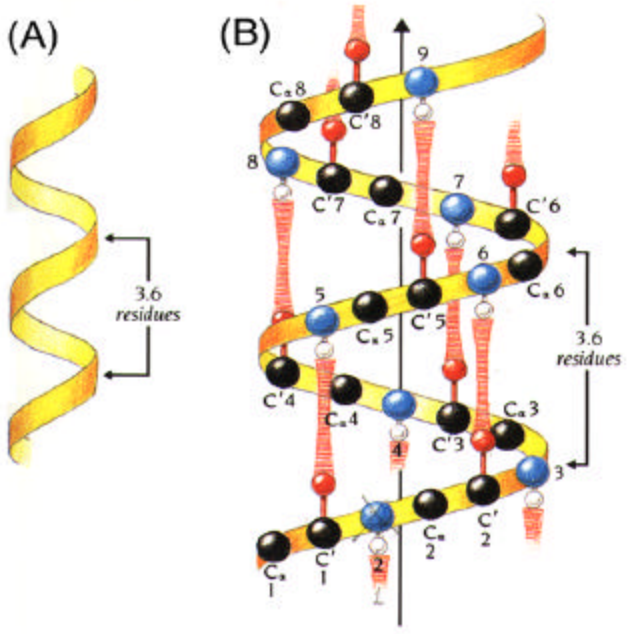


Figure II.3.2. *The a-helix conformation in proteins. (A) idealized drawing of the path of the main chain in an a-helix, with 3.6 residues per turn, which corresponds to 5.4 Å distance between successive turns (or 1.5 Å per residue). (B) includes the approximate positions*

of the main chain atoms. The atoms are labeled by the residue they belong to, starting from the N-terminus. Black balls are carbon atoms, blue ones along the backbone are nitrogens, and those red ones appended to the backbone carbon atoms are the oxygen atoms. Amide hydrogens are shown by white balls. Side chains are not shown. Springlike connections between the carbonyl O of residue i and the amide H of residue i+4 indicate the (i, i+4) hydrogen bonds. (from (Branden and Tooze, 1999))

There are four factors that can contribute to the stability of an α-helix: the hydrogen bonds, the conformational preferences for the helical (f, y ? angles - in the α-helix (see Ramachandran plots in § II.2); the side chains do not interfere with the backbone in those conformations; the favorable van der Waals interactions inside the helix due to the small hole down the helix axis; and the good alignment of the electrical dipole moments of the amino acids parallel to the helix axis. *Figure II.3.3* shows that sidechains reside on the outsides of helices. Like the individual amino acids themselves, helices have chirality. Almost all helices in proteins are *right-handed*. In this way, the chain avoids steric conflicts between carbonyl groups and the side chains of *L*-amino acids, as discussed in § II.2.c. But, there is an effect of the helix on side chain conformations (see *Fig. II.2.9(B)*), where a gauche$^-$ state for χ_1 is excluded by steric conflicts with the helical backbone. If proteins were made of D-amino acids, then helices would be left-handed.

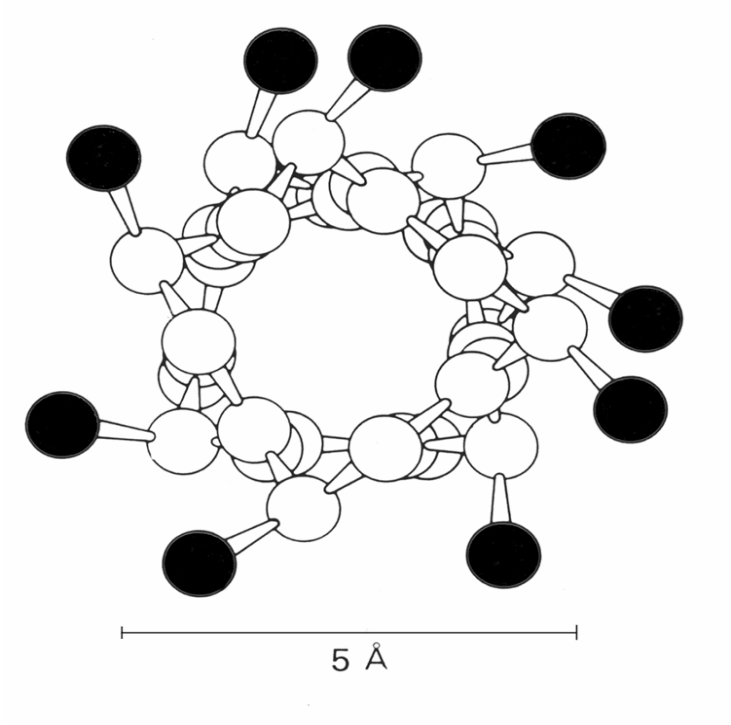
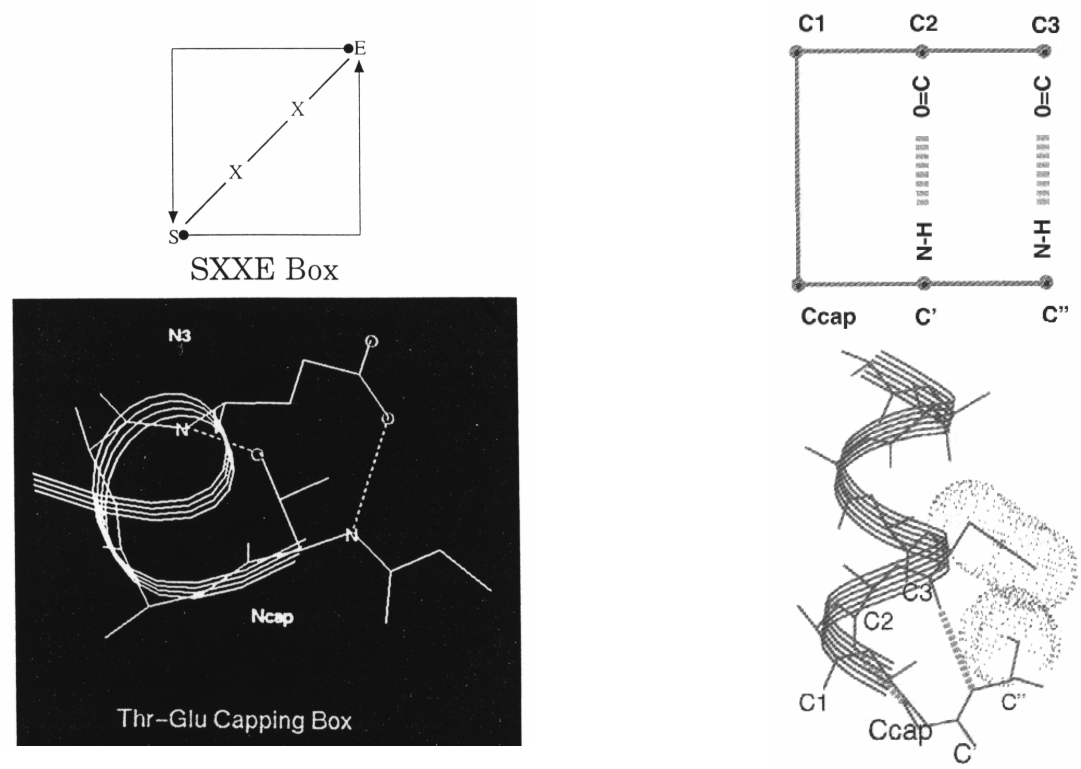


Figure II.3.3. *Cross-sectional view of an a-helix. The sidechains (shaded spheres) are on the outside. The van der Waals radii of the atoms are larger than shown in the ball-and-stick model here. There is almost no free space inside of an a helix. (from (Stryer, 1988))*

At the ends of helices, carbonyl and amide groups will have unsatisfied hydrogen bonds. In such cases, there are often *end caps*, where a side chain that can form a hydrogen bond folds back onto the backbone to replace the unfulfilled backbone hydrogen bonds (see *Figure II.3.4*).



A **B**

Figure II.3.4. *Helix capping interactions in proteins. Helical residues are labeled as ---N"-N'-N_{cap}-N1-N2-N3-…-C3-C2-C1-C_{cap}-C'-C"-… where N1 through C1 are residues with helical ƒ,y values.* ***(A)*** *The N-terminus capping box is a hydrogen-bonded cycle in which the sidechain of N_{cap} forms a hydrogen bond with the backbone amide hydrogen of N3, and reciprocally, the sidechain of N3 forms a hydrogen bond with the backbone amide hydrogen of N_{cap}. (Top) Ser-X-X-Glu capping (Harper and Rose, 1993)) where each X denotes any amino acid. The diagonal line represents the covalent connectivity; the hydrogen bonds are represented by the arrows on both sides of the box. (Bottom) Thr-X-X-Glu capping box. The helix is represented by the ribbon; residues N_{cap} and N3 shown in atomic detail. Hydrogen bonds are represented by dashed lines (from (Harper and Rose, 1993))* ***(B)*** *One of the C-capping motifs from cytochrome C551. This conformation belongs to a class called glycine-terminated helices because C' is a glycine. (Top) The backbone-backbone hydrogen bonding pattern (dashed lines). (Bottom) Helix terminating in a "Schellman motif". The stippled surfaces show the van der Waals interactions between the sidechains of C" (Cys) and C3 (Phe). Dashed lines indicate the hydrogen bonds from C" to C3 and C' to C2. (from (Aurora et al., 1994))*

A *helical wheel diagram is* a 2-dimensional projection of a helix onto a plane perpendicular to its axis. It shows the periodicity of amino acids around the helix; see Figure *II.3.5.* What is often observed in such diagrams is that for helices in globular proteins, the hydrophobic amino acids tend to cluster on one side of the helix (pointing toward the interior of the protein), and the polar and charged amino acids are on the outer face. These are called *amphipathic* helices, and are said to have a *hydrophobic moment (Eisenberg and McLachlan, 1986)*.

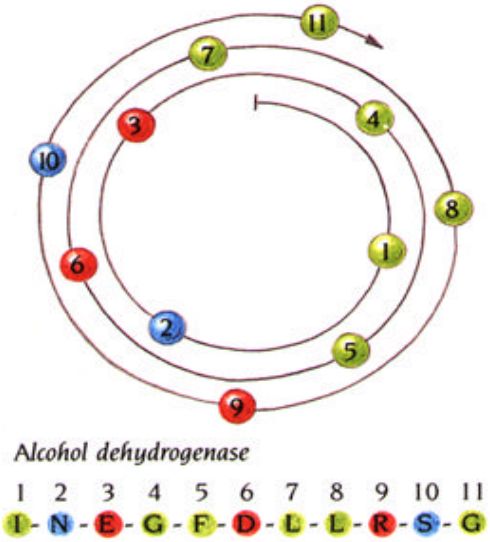


Figure II.3.5. *Helical wheel diagram. Following the α-helix geometry, residues along the given sequence are plotted successively at 100° intervals. The above sequence correspond to an amphipathic helix in the enzyme alcohol dihydrogenase. Hydrophobic and polar/charged sidechains are represented by green and red or blue balls, respectively. (from (Branden and Tooze, 1999))*

b-sheets. When Pauling and his colleagues predicted the α-helix, they also predicted that hydrogen bonding would lead to parallel and anti-parallel traintrack-like structures they called "β-sheets" (see *Figs. II.3.6* and *7*). A β-sheet is comprised of individual *strands* each of which is a nearly planar zigzag. Strands pair through amide-to-carbonyl hydrogen bond links. The

side chains lie above or below the sheet, and they are well-placed to interact with neighboring side chains. β-sheets are stabilized by hydrogen bonds, by their side chains' interactions, by favorable (*f*, *y* ?angles (in the β-region of the Ramachandran map), and by van der Waals attractions. *Figure II.3.8* shows that large β-sheets are not planar; but actually have a twist. The regularity of β-sheets can sometimes be interrupted by a 'β-bulge'. An example is shown in *Figure II.3.9.*

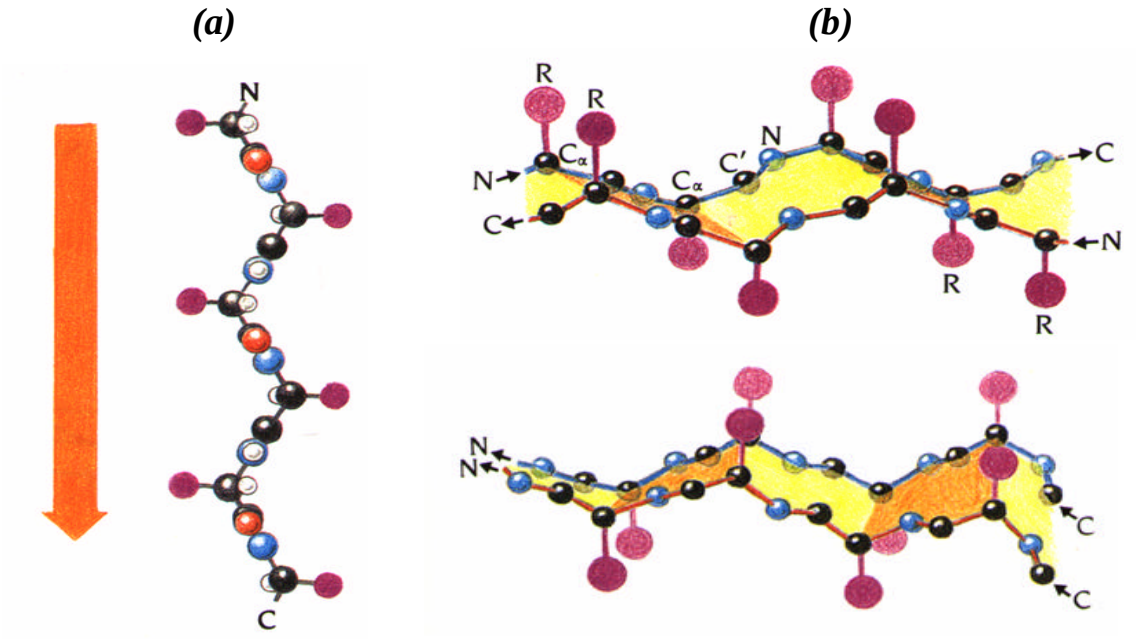


Figure II.3.6. *Schematic illustration of b-sheets.* ***(a)*** *The extended conformation of a b-strand in a ball-and-stick model. The arrow indicates the direction from the N-terminus to the C-terminus. Sidechains are shown as purple spheres, N atoms are colored blue, and carbonyl carbons, red.* ***(b)*** *Illustration of the pleat of b-sheets for two antiparallel strands (top) and two parallel strands (bottom). (adapted from Figures 2.5(a), 2.5(d) and 2.6(c) of (Branden and Tooze, 1999)*

(a) *(b)*

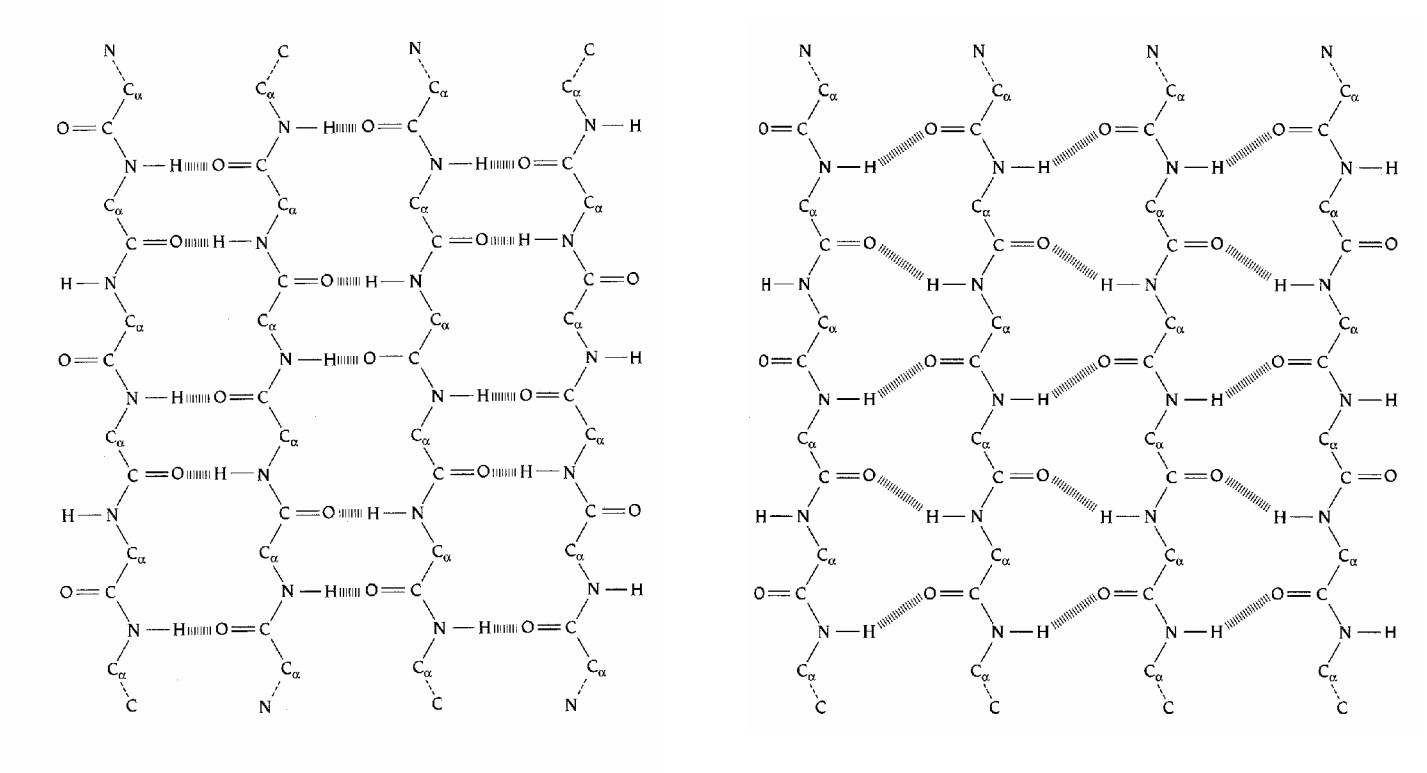


Figure II.3.7. *Hydrogen bonding pattern in (a) antiparallel and (b) parallel b-sheets. (adapted from Figures 2.5(b) and 2.6(b) of (Branden and Tooze, 1999).*

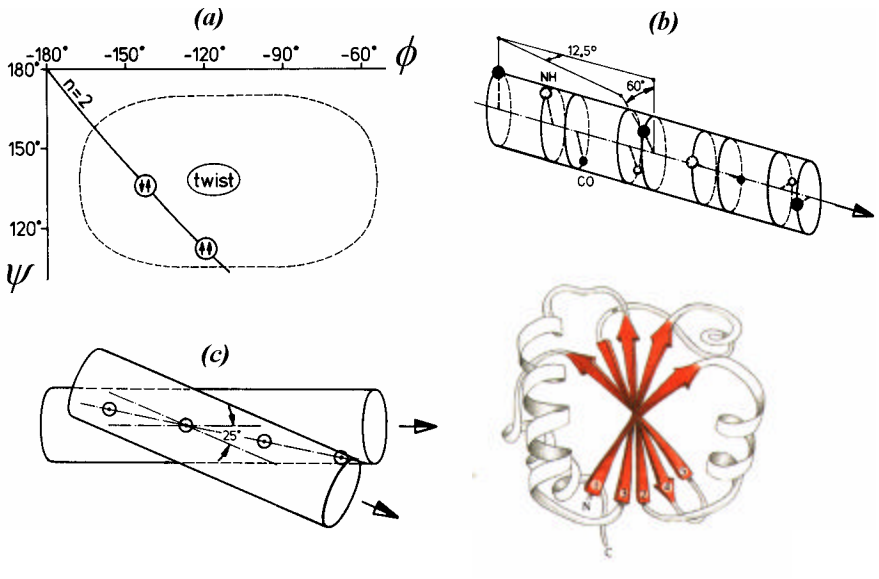


Figure II.3.8. *Twist of b-pleated sheets.* ***(a)*** *Region of (f,y ?map corresponding to the b-sheet region (region II in Figure II.2.6(a)). The diagonal indicates the loci of dihedral angles in planar zigzag (2-fold helical) structures. The dihedral angle positions of the ideal parallel (- -) and antiparallel (- ¯) b-sheets are indicated. The dashed contour encloses the allowed b-region (having energy lower than 1 kcal/mol). Note that the center of this region (label "twist") is to the right of the diagonal. This corresponds to (f, y) values of typical strands in twisted sheets.* ***(b)*** *Right-handed twist along a single strand in a twisted sheet. The rotations of the hydrogen-bonding directions of the carbonyl (filled circles) and amide (open circles) groups are indicated. Twist angles of b-strands vary considerably. Angles given here are only typical values.* ***(c)*** *shows how two parallel strands twisted as sketched in part (b) can pair up to form hydrogen bonds if the backbone directions of the two strands make an angle with each other. This leads to the left-handed twist characteristic of b-sheets observed in a number of proteins.* ***(d)*** *Ribbon diagram of the protein thioredoxin from E coli, illustrating the left-handed twist of a b-sheet. (parts (a)-(c) adapted from Figure 5.10 in (Schulz and Schirmer, 1979), and (d) taken from Figure 2.7(a) of (Branden and Tooze, 1999))*

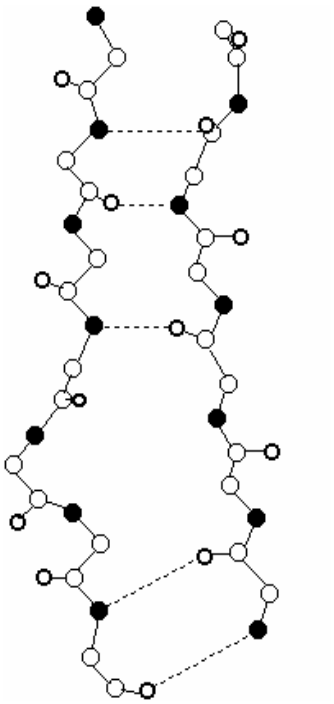


Figure II.3.9. *A b-bulge in a sheet in immunoglobulin fragment V_L. (taken from (Lesk, 1991))*

Other secondary/regular structures. Also possible, but much rarer than the α-helix, are the 3_{10}- and π-helices, where hydrogen bonds are formed between (i, i + 3) and (i, i + 5), respectively. The 3_{10}- helix is so-named because it involves 3 residues per turn and 10 atoms in the ring enclosed by the hydrogen bond. The 3_{10}- helix is sometimes found in short peptides and occasionally in proteins because although its hydrogen bonds are less stable than those in the α-helix (side chain packing is less favorable, dipoles are more poorly aligned), there are more hydrogen bonds. π-helices are quite rare because of *f y* steric hindrance and because of a potentially large hole down the helix axis, which would result in a loss of van der Waals interactions. See *Table II.3.1* for the details of the geometry of these secondary structures.

Table II.3.1.
Approximate Geometric Parameters for Some Regular Protein Conformations

Secondary structure	Dihedral Angle f	y	w	Residues[a] per turn & chirality	Translation per residue (Å)
Right-handed α-helix	-57	-47	+180	+3.6	1.50
Left-handed α-helix	+57	+47	+180	-3.6	1.50
3_{10}-helix	-49	-26	+180	+3.0	2.00
π-helix	-57	-70	+180	+4.4	1.15
Antiparallel β-sheet[b]	-139	+135	-178	±2.0	3.40
Parallel β-sheet[c]	-119	+113	+180	±2.0	3.20
Twisted antiparallel or parallel β-sheet[d]	≈-110	≈+140	≈+180	≈-2.4	≈3.3
Fully extended chain[e]	-180	+180	+180	±2.0	3.6

Partly adapted from *Table 2* of {IUPAC-IUB Commission on Biochemical Nomenclature. 1970. Abbreviations and symbols for the description of the conformation of polypeptide chains. Tentative rules (1969) [published simultaneously in Biochemistry 9: 3271-3479, J. Biol. Chem. 24: 6489-6497; J. Mol. Biol. 52: 1-17]} and *Table 5.1* of (Schulz and Schirmer, 1979).

[a] + and – correspond, respectively to situations when successive C^{α}-C^{α} virtual bonds follow a right-handed and a left-handed helical path, ± is used when such helices become planar, i.e. when the C^{α}-C^{α} virtual bonds follow a planar zigzag form.

[b, c] Pleated sheets with Pauling-Corey idealized geometries, rare in proteins.

[d] Twisted β -sheets are abundant in proteins. There are considerable variations among the dihedral angles of twisted β-strands. The values given here only roughly indicate the location of the center of the β -sheet region on the (ϕ,ψ?map; see *Figure II.3.8*.

[e] Included for reference only. Fully extended chains are not part of secondary structures. They do not form stable sheet-like chain organizations.

Turns and loops. The secondary structures described above - α-helices and β-sheets - are regular and repeating. When there is a reversal of chain direction, secondary structures are connected together by *turns* (also called *reverse turns*) and *loops.* See *Figure II.3.10.* According to the original definition of Linderstrom-Lang in 1953, secondary structure meant only helices. Soon thereafter, "secondary structure" came to include sheets. We follow that convention in this text: and refer to secondary structures as only the repeating regular structures - helices and sheets, but not turns and loops. Turns are usually short and tight, in the range of 3-5 monomers, typically self-hydrogen bonded; loops are longer.

Reverse turns occur almost exclusively at the protein surfaces. Reverse turns therefore contain polar residues, together with glycine and proline. As shown in § II.2, having no side chain hindrance, Gly can adopt a broad range of dihedral angles giving rise to kinks in the main chain. Pro is unique, because its sidechain curls back to the main chain and seizes it.

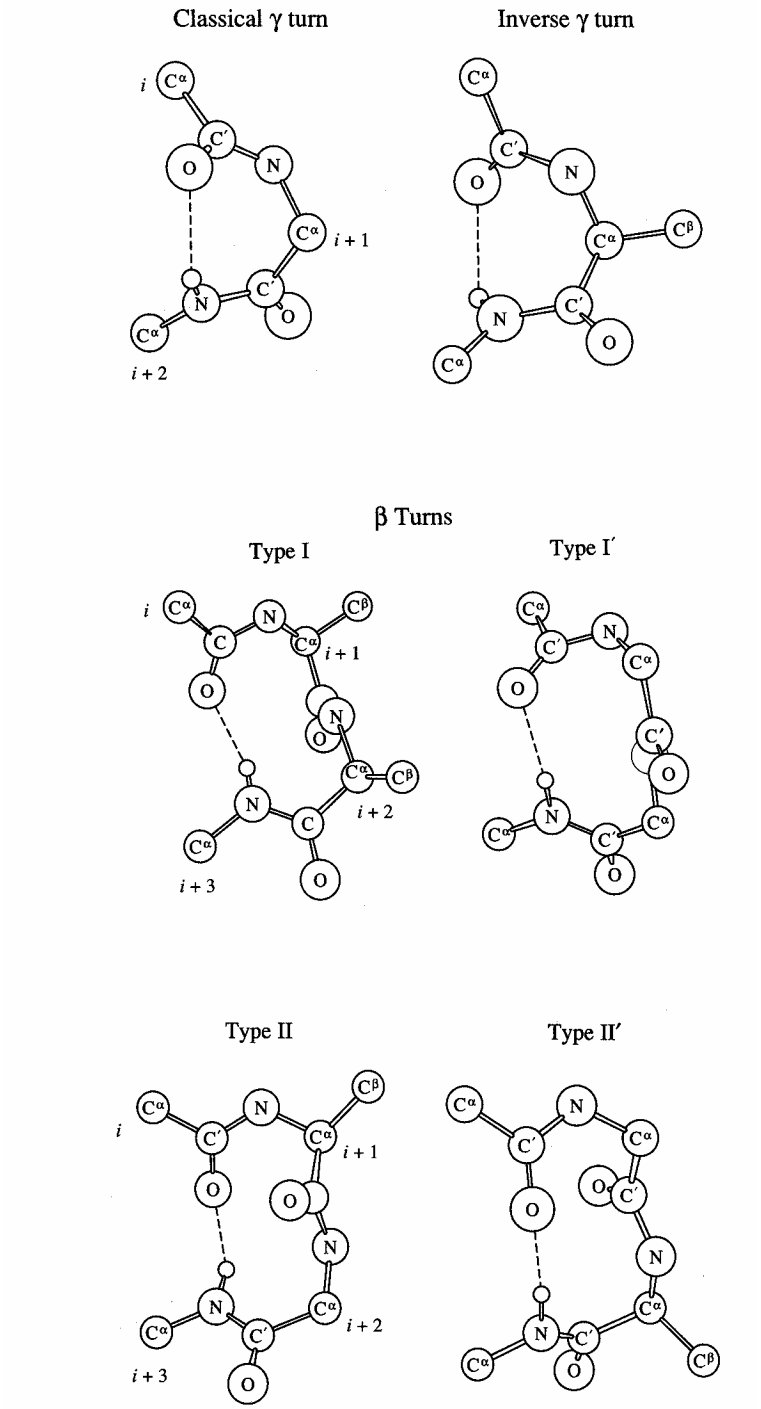


Figure II.3.10. *Common reverse turns g and b turns connect adjacent strands of an antiparallel b-sheet. The dashed lines indicate the last hydrogen bond of the b-sheet. g?turns are very tight; they are made up of three residues, one of which is not involved in hydrogen bonding. The more common b turns involve four residues, two of which do not form hydrogen bonds. Types I' and II' backbone conformations are mirror images of types*

I and II backbone conformations, respectivel. .C^b atoms are included only at positions where residues other than Gly occur frequently. Not shown here are Type III b turns. They may be considered as very short segments of 3_{10}-helical conformation. Many other types of turns have been defined as well. (taken from (Creighton, 1993))