Proteins: Fundamental Chemical Properties

Introductory article

Alain J Cozzone, *Institute of Biology and Chemistry of Proteins, CNRS, Lyon, France*

Article Contents

Proteins are important biological polymers formed from building blocks called amino acids. The three-dimensional structure and biological activity of proteins depend on the physicochemical properties of their constituent amino acids.

Introduction

Proteins are macromolecules found in all biological systems, from lower prokaryotes to higher eukaryotes. They occupy a prominent position in living cells, both quantitatively and qualitatively, which accounts for the origin of their name derived from the Greek word *prôtos*, meaning 'first rank of importance'.

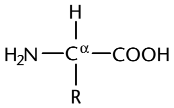
Quantitatively, proteins are the most abundant class of biomolecules since they represent over 50% of the dry weight of cells, far more than other important biopolymers such as nucleic acids, polysaccharides or lipid assemblies. Each organism contains a large variety of specific proteins, according to the number of the corresponding genes present in chromosomes. This number varies from a few hundreds in certain bacterial species to several thousands in animals and man.

Qualitatively, proteins are involved in virtually all biological processes. Thus, most chemical reactions occurring in life forms are catalysed by specific proteins called enzymes that are able to increase reaction rates by several orders of magnitude. Proteins can also transport and store a wide array of ions and small molecules as well as electrons. They possess hormonal activity and, in the form of antibodies which distinguish between self and nonself, they defend organisms against intruders. They participate in the reception and transmission of signals and stimuli at both intracellular and intercellular levels. They play crucial roles in the regulation of the expression of genetic information, at either transcription or translation, in connection with the control of growth and differentiation of cells. They are also necessary for providing the mechanical support and filamentous architecture within and between cells, and consequently are essential to cellular contraction and coordinated motion.

Whatever their structural or functional role, all proteins are polymers composed of the same building blocks, the amino acids, which are covalently joined together by amide links, known as peptide bonds. They differ only in the number, the nature, and the sequential order of their constituent amino acids. To understand the functional diversity of proteins, it is important, first, to appreciate the physicochemical properties of the different amino acids, even though the properties of a protein molecule are hugely more complex than the sum of the properties of its different constituent amino acids. It is then possible to determine the three-dimensional structures that these linked building blocks can acquire and to analyse the biological properties of the corresponding polymers.

Chemical Composition and Properties of Proteins

Proteins from all organisms (viruses, bacteria, plants, animals, etc.) are constituted from the same 20 different amino acids (**Table 1**). Each amino acid comprises an amino group, a carboxyl group, a hydrogen atom, and a specific R group bonded to a carbon atom called the α-carbon [**I**]. The R group is referred to as the 'side-chain' and varies in size, charge, shape and chemical composition from one amino acid to the other.



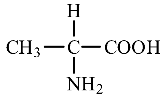
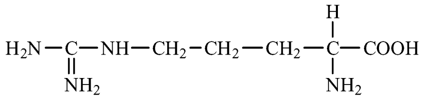
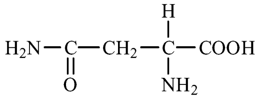
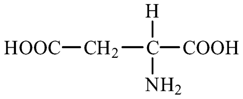
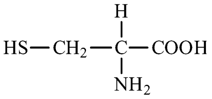
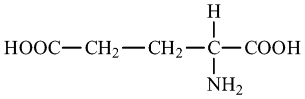
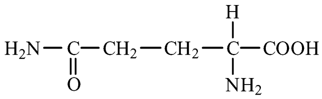
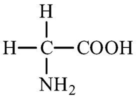
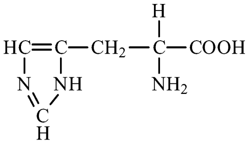
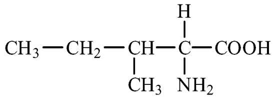
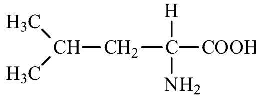
[I]

Characteristics of amino acids

Of the 20 amino acids usually present in proteins, 19 have the general structure shown in [**I**], but one amino acid – proline – has its side-chain bonded to the nitrogen atom to give an imino acid. The additional infrequent occurrence of a twenty-first amino acid, termed selenocysteine, has been reported. In this molecule, the sulfur atom of cysteine is replaced by selenium.

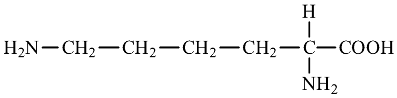
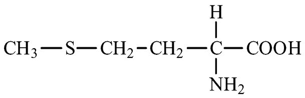
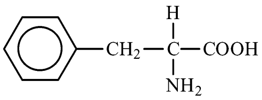
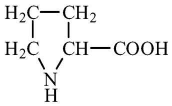
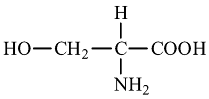
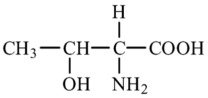
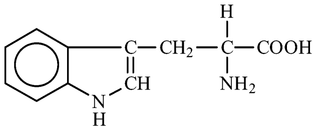
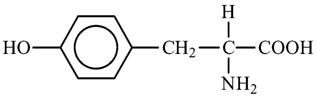
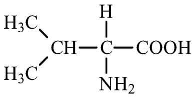
In all the amino acids except glycine, where the side-chain is just a hydrogen atom, the α-carbon is asymmetric, but of the two possible optical isomers, D and L, only the L-isomer is found in natural proteins. Two amino acids, threonine and leucine, possess an additional asymmetric

Table 1 List of 20 fundamental amino acids and their abbreviations

Amino acid	Three-letter abbreviation	One-letter symbol	Formula
Alanine	Ala	A	
Arginine	Arg	R	
Asparagine	Asn	N	
Aspartic acid	Asp	D	
Cysteine	Cys	C	
Glutamic acid	Glu	E	
Glutamine	Gln	Q	
Glycine	Gly	G	
Histidine	His	H	
Isoleucine	Ile	I	
Leucine	Leu	L	

continued

Table 1 – *continued*

Amino acid	Three-letter abbreviation	One-letter symbol	Formula
Lysine	Lys	K	
Methionine	Met	M	
Phenylalanine	Phe	F	
Proline	Pro	P	
Serine	Ser	S	
Threonine	Thr	T	
Tryptophan	Trp	W	
Tyrosine	Tyr	Y	
Valine	Val	V	

centre in their side-chains and, of the four possible optical isomers, only one is used biologically.

The carboxyl and amino groups of amino acids are ionized in solution at neutral physiological pH, with the carboxyl group bearing a negative charge ($-COO^-$) and the amino group a positive charge ($-NH_3^+$). The ionization state varies with pH: in acid solution the carboxyl group is not ionized ($-COOH$) and the amino group is ionized ($-NH_3^+$). Conversely, in alkaline solution the carboxyl group is negatively charged ($-COO^-$) and the amino group is not ionized ($-NH_2$).

Classification of amino acids

In general, the ionization properties of the constituent amino acids, including those of their side-chains, greatly influence the solubility, stability and structural organization of proteins. Similarly, the hydrophobicity/hydrophilicity of the side-chains plays an important role in the physicochemical behaviour of polypeptide chains and their folding into three-dimensional structures. Different classifications of the 20 amino acids have been proposed on the basis of the properties of their R groups. One of the most commonly used includes six categories:

1. Aliphatic (5 amino acids): glycine, the simplest amino acid with only a hydrogen atom in its side-chain; alanine, valine, leucine and isoleucine, which contain noncyclic hydrophobic nonpolar chains, poorly soluble in water.
2. Hydroxylic (2 amino acids): serine and threonine which contain, respectively, a primary and a secondary alcohol group, are polar and very soluble in water.
3. Acidic or dicarboxylic, and corresponding amides (4 amino acids): aspartic acid and glutamic acid, which both possess a second carboxyl group in their side-chain that is ionized, negatively charged and very polar under physiological conditions; the corresponding derivatives, asparagine and glutamine, which contain a polar amide group in place of the second carboxyl group.
4. Basic (3 amino acids): lysine and arginine, which are ionized, positively charged and very polar, under most physiological conditions; histidine, with an imidazole group, which is poorly protonated at pH 7.
5. Cyclic (4 amino acids): phenylalanine, tyrosine and tryptophan, which each contain an aromatic cycle; proline, the imino acid with an aliphatic heterocycle.
6. Sulfur-containing (2 amino acids): methionine, with a hydrophobic nonpolar chain; and cysteine, which is polar because of its sulfhydryl group.

Soluble proteins and membrane proteins

Proteins exist essentially in either aqueous or membrane environments. The presence of polar groups in the side-chains of amino acids located at the surface of proteins, unlike hydrophobic groups which are packed inside, favours solubility of the proteins in water. This solubility is affected mainly by the addition of salts. Thus, at low ionic strength, the solubility of most proteins is relatively high (salting-in effect), but it is reduced when the ionic strength increases (salting-out effect).

Besides the ionic strength, the solubility of proteins is also affected by the pH value: it is minimal at pH values close to their isoelectric point where the net charge is equal to zero. Solubility is also dependent on the temperature and the dielectric constant of the solvent. In addition, when any protein is unfolded or partially denatured, it is generally less soluble than in its folded form because the nonpolar groups are then more exposed.

In general, proteins are insoluble in nonpolar solvents and the extent of their solubility is determined by the interactions of their polar groups with water. Thus, membrane proteins have, on the one hand, a polar surface that interacts with water or aqueous solutions and with the lipid head groups and, on the other hand, a nonpolar surface that interacts with the nonpolar interior tail of the basic structure of membranes, the lipid bilayer. Therefore, their solubility is low both in aqueous media and nonpolar solvents.

The extent of association of proteins with membranes *in situ* is related to the length of the polypeptide chain that is immersed in the membrane bilayer. The integral membrane proteins are almost completely immersed in the bilayer with only their two ends in contact with aqueous solvents. The nonintegral membrane proteins are, by contrast, much more exposed to those solvents and are mostly water soluble.

Detection of amino acids and proteins

The detection of amino acids and proteins is achieved by physical, chemical and biological techniques. The presence of amino acids, when polymerized, can be revealed by the ultraviolet absorbance of the peptide bond below 230 nm. When they are either free or linked in polymers, the presence of amino acids can also be detected by their absorbance around 280 nm for those bearing an indole ring (Trp) or an aromatic ring (Phe, Tyr).

Since proteins contain about 16% nitrogen by mass, they can be assayed by the Kjeldahl method based on sulfuric acid mineralization and measurement of the released ammonia. Another procedure is the biuret reaction that specifically measures peptide bonds: copper(-II) sulfate in alkaline tartrate reacts with a peptide bond to produce a purple compound with maximum absorption at 540 nm. A similar, and more sensitive, assay is the Folin phenol method developed by Lowry and collaborators, in which copper ions are added to the protein solution in the presence of a phosphomolybdate–tungstate mixture. Amino groups of amino acids and proteins also react with ninhydrin to give a purple product with maximum absorbance at 570 nm. Fluorescamine and *o*-phthalaldehyde also react with amino groups and form fluorescent products.

A common reagent for staining proteins is Coomassie brilliant blue which can be used to quantify proteins in solution or included in gels. Another sensitive staining method involves silver nitrate. Biological assays take advantage of the immunological properties of proteins, their affinity for specific ligands or their enzymatic activity.

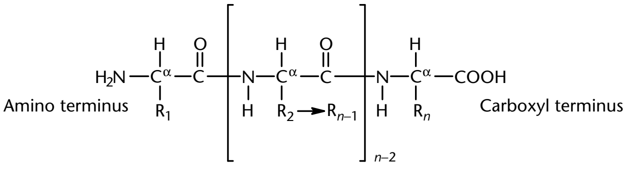
Primary Structure

The primary level of structure refers to the linear sequence of amino acids along a protein chain and to the location of covalent bonds, namely disulfide bonds, between chains or within a chain. The primary structure identifies a protein unambiguously, determines its chemical and biological characteristics, and specifies the higher levels of protein structure.

Formation of the polypeptide chain

During protein synthesis, the amino acids are joined end-to-end through covalent linkage. The α-carboxyl group of the first amino acid reacts with the α-amino group of the next amino acid to generate a peptide bond and eliminate a water molecule. The equilibrium of this reaction lies on the side of hydrolysis rather than synthesis. Therefore, the process of protein synthesis requires a supply of free energy.

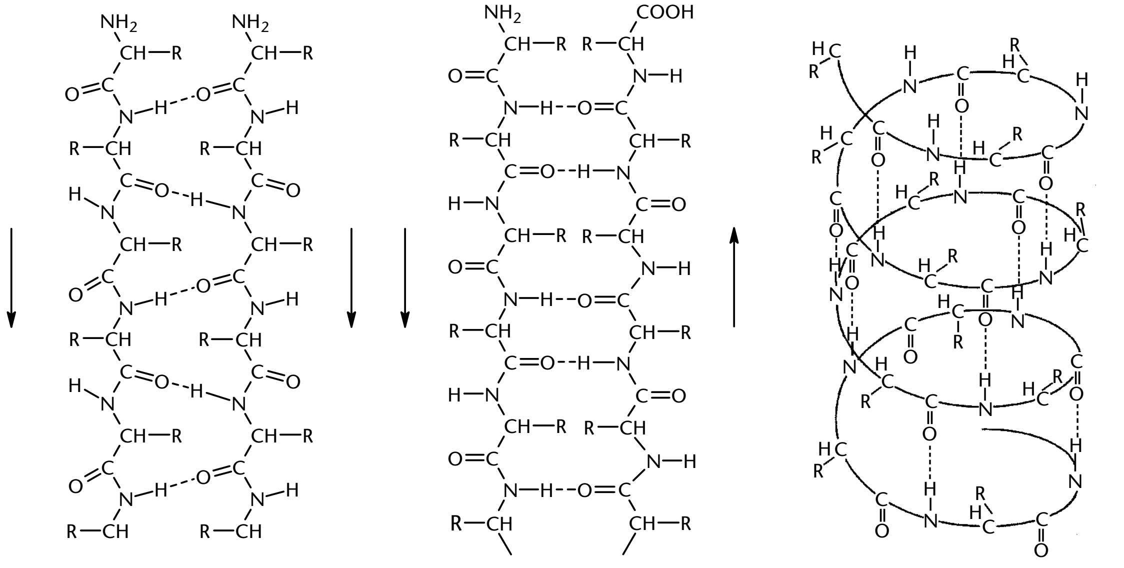
A consequence of this process is that the amino group of the first amino acid of a polypeptide chain and the carboxyl group of the last amino acid remain intact. A polypeptide chain has thus two different ends, termed respectively the N-terminus and the C-terminus. Amino acids are polymerized from the N-terminus to the C-terminus. They are numbered in that direction, e.g. Arg18 represents an arginine at position 18 from the amino terminus, and the generally accepted convention is to write the amino acid sequence of a polypeptide from left to right as shown in [**II**].



[**II**]

A polypeptide chain consists of a regularly repeating part, called the 'main chain' or 'backbone', from which projects a variable part comprising the different side-chains. Each amino acid unit in a protein is called an amino acid 'residue' whose specific name derives from the corresponding free amino acid, e.g. arginyl residue (arginine), lysyl residue (lysine), and so on.

The peptide bond –CO–NH– between two successive residues is a relatively rigid planar structure because of its partial double-bond character due to resonance. As a consequence, rotation of this bond is restricted. However, there is rotational freedom about the single bonds that link each C^{α} atom to the N and C atoms of peptide bonds. The convention is to denote the angle of rotation around the N–C^{α} bond as phi (Φ) and to denote the angle around the C^{α}–C bond from the C^{α}-atom as psi (Ψ). The planar peptide bond can theoretically exist in two different configurations, the *trans* and *cis* forms [III], but the *trans*



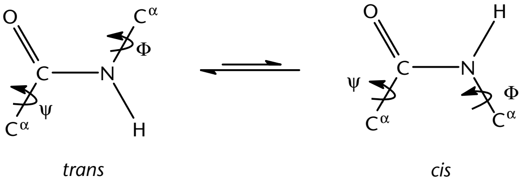
Parallel β pleated sheet

Antiparallel β pleated sheet

Right-handed α helix

Figure 1 Major periodic elements of protein secondary structure.

form is energetically favoured over the *cis* form.



[III]

Determination of the primary structure

The first determination of the complete amino acid sequence of a protein was achieved in 1953 by Frederick Sanger in the case of insulin, a protein hormone of 51 residues. Since then the sequences of several hundred proteins have been deciphered using various techniques based on specific chemical reactions and special procedures for separating and identifying peptides and amino acids.

Although it is currently easier to determine the sequences of proteins from their gene sequences, protein sequencing techniques are still required in designing oligonucleotide probes for molecular genetic experiments. They are also necessary for assessing the various modifications of proteins occurring during or after biosynthesis (see below), which are fundamental for acquiring proper structure and biological activity.

A first approach to sequence determination consists in identifying the N-terminal and C-terminal residues. The N-terminal residue can be identified by several chemical techniques based on specific labelling with a compound, such as 1-fluoro-2,4-dinitrobenzene or dansyl chloride or cyanate, with which it forms a stable covalent link. Similarly, the C-terminal residue can be identified by hydrazinolysis after conversion of the α-carboxyl groups of peptide bonds into hydrazides. Alternatively, these chemical procedures can be replaced with enzymatic techniques based on the sequential release of single amino acids from the ends of chains using specific exopeptidases, namely aminopeptidase and carboxypeptidase.

One of the most efficient procedures for determining the amino acid sequence of a protein is the Edman degradation method. In this case, amino acids are removed sequentially, one at a time, from the N-terminus without cleaving the peptide bonds between the other amino acid residues of the chain. The terminal α-amino group is first reacted with phenylisothiocyanate in alkaline medium to yield the phenylthiocarbamoyl derivative; this derivative is then released from the rest of the chain, by acid treatment, in the form of a cyclic compound that rearranges in aqueous solution to the phenylthiohydantoin derivative, which is further identified by chromatographic techniques. The procedure can be repeated as many times as necessary to determine the complete amino acid sequence of the protein. Such recurrent analysis is usually performed in an automatic amino acid sequencer.

In many instances, namely for analysing larger proteins, it is useful to cleave the polypeptide chain into smaller peptides before sequencing by the Edman method. Specific cleavage can be achieved either by chemicals (e.g. cyanogen bromide, hydroxylamine, *N*-chlorosuccinimide) or by enzymes (e.g. trypsin, pepsin, elastase, thermolysin).

An alternative technique for determining the primary structure of proteins is mass spectrometry. Basically, the protein is rendered volatile by chemical treatment of side-chains and fragmented nonspecifically with an electron beam. The protein fragments are separated according to their mass-to-charge ratio and identified. Chemical modification of side-chains can be avoided in more recent techniques and replaced by ionization of proteins with a high-energy beam of atoms or ions. Otherwise, a solution of protein in a volatile solvent can be sprayed directly into the mass spectrometer (electrospray ionization). This technique is very sensitive (picomole range) and accurate.

Covalent Modifications of Proteins

During or after biosynthesis of the polypeptide chain, a number of covalent modifications often occur. One of these consists of the proteolytic cleavage of one or several residues. The process concerns, for instance, the removal of signal peptides that target proteins to various locations within the cell. It also applies to a series of proteins, those destined for cellular organelles or for secretion, that are initially synthesized in the form of larger precursors (zymogens, prohormones, proproteins, preproproteins) from which they are released by proteolysis.

Another example concerns polypeptides that are modified after translation by excision of internal segments (inteins) through a self-splicing mechanism. In addition, the side-chains of numerous amino acid residues, as well as the N-and C-termini, can be covalently modified by enzymatic addition of various chemical groups that generally affect the biological properties of the protein.

About 200 modifications of this type, reversible or irreversible, have been detected so far, some of which are extremely rare while others are very frequent. The latter include the binding of lipid polar groups to proteins, at either their N-terminus (myristoyl groups) or their C-terminus (glycosyl-phosphatidylinositol and farnesyl groups); glycosylation of either *N*-type (on the nitrogen atom of asparaginyl side-chains) or *O*-type (on the oxygen atoms of hydroxyls, particularly those of seryl and threonyl residues); phosphorylation, usually at the hydroxyl groups of seryl, threonyl and tyrosyl residues in a wide array of proteins belonging to all biological systems, and also at aspartyl and histidyl residues in the particular case of bacterial proteins; hydroxylation at prolyl and lysyl residues, which is crucial for the maturation and secretion of proteins such as collagens; methylation; acetylation; sulfation; carboxylation; ADP-ribosylation; nucleotidylation; and amidation.

Some proteins, called heteroproteins, contain nonpeptidic prosthetic elements, covalently or noncovalently bound to them, that are essential to their activity: haems, porphyrins, nucleotides, metal ions, etc.

Secondary Structure

In principle, a polypeptide chain could assume great flexibility owing to the free rotation of the atoms around different bonds along the chain. If it did so, it would behave like a random coil and could theoretically adopt a myriad of conformations of similar energies. In fact, in biological conditions, each protein adopts essentially only one conformation because the side-chains of its amino acid residues associate locally with one another and with the solvent to yield a global structure of maximum stability. The correct folding of proteins sometimes requires the assistance of a particular class of cellular proteins called molecular chaperones.

The interactions involved are mostly of a noncovalent type: hydrogen bonds where a hydrogen atom covalently bonded to an electronegative atom (the donor) interacts favourably with another electronegative atom (the acceptor); electrostatic forces between net charges of opposite signs, known as salt bridges, or between two dipoles each formed from the asymmetric distribution of electrons within the covalent linkage of two atoms having different electronegativities; repulsion forces between electron orbitals of atoms, which operate like impenetrable spheres defined by their van der Waals radius; and hydrophobic interactions between nonpolar residues, which give the largest single contribution to protein stability. Besides these noncovalent interactions, the conformation of proteins is also stabilized by some covalent bonds, especially disulfide bridges between pairs of cysteyl residues.

Several folding patterns occur repeatedly in parts of protein molecules. They are known collectively as secondary structure, which constitutes the next level of protein structure after the primary structure. These regular arrangements of the linear polypeptide chains with repeating values of the Φ and ψ torsions angles and main-chain hydrogen bonding are of two major types: α helices, with repeating patterns of local hydrogen bonding, and β sheets, with repeating patterns between distant parts of the polypeptide chain.

The α helix structure

The α helix is a right-handed rodlike structure,which means turning in a clockwise direction as viewed from the near end, or else turning in the direction of the fingers of the right hand when the thumb indicates the line of sight. Its inner part is formed by the coiled polypeptide main chain and the surface by the side-chains projecting outwards in a helical arrangement (**Figure 1**). It is stabilized by hydrogen bonds between the C = O group of each amino acid residue of the main chain and the N–H group of the residue located four residues away in the amino acid sequence, with repeated torsion angle values of about $-60°$ for Φ and $-40°$ for ψ. It is also stabilized by the van der Waals interactions generated by the close packing of the backbone atoms.

A single turn of the helix contains 3.6 residues and, since a residue extends 1.5 Å, the pitch of the helix is 5.4 Å (1.5×3.6). Along the helix, all the hydrogen bonds and peptide groups point in the same direction, nearly parallel to the helix axis. Since each peptide bond possesses an individual dipole moment, the overall effect is a cumulative macrodipole for the helix with a positive charge at the amino end and a negative charge at the carboxyl end.

The α helix structure was first predicted by Linus Pauling and Robert Corey in 1951 on the basis of crystallographic analyses of the structures of various small molecules. This prediction was strongly supported, soon after, by diffraction patterns obtained by Max Perutz from haemoglobin crystals and keratin fibres. It was fully demonstrated by John Kendrew from the X-ray reconstruction of the structure of myoglobin, whose secondary structure consists exclusively of α helices (eight in total).

The lengths of α helices in natural proteins vary from a few to several tens of residues, with an average of about 10 residues. The α helix content varies widely from one protein to another; it is very high in some globular proteins such as haemoglobin, myoglobin and ferritin, or in fibrous proteins such as α-keratin, myosin, epidermin and fibrinogen; by contrast, it is relatively low or zero in other proteins such as chymotrypsin, superoxide dismutase and cytochrome *c* (see below). The various amino acid residues along a polypeptide chain have different tendencies to form α helices. For example, Ala, Leu, Phe, Trp, Met, His and Gln stabilize α helices, whereas Ser, Ile, Thr, Gln, Asp and Gly have a destabilizing effect and, even more so, proline and hydroxyproline create sharp bends in the helices that destroy them.

The α-helix most frequently present in proteins is the right-handed helix, described above. A left-handed α helix is also possible sterically, but the side-chains are too close to the main chain and, therefore, this conformation is unstable and rarely encountered in natural polypeptides. Other types of α helix can be envisaged, with hydrogen bonds between residues nearer together ($n + 3$) or farther apart ($n + 5$). The former is called the 3_{10} helix, with 3 residues per turn and 10 atoms between the donor and the acceptor in the hydrogen bond; it is rarely found, except at the end of α helices. The latter is designated the π helix; it has never been observed in proteins.

The β sheet structure

The other major type of periodic secondary structure is called the β sheet (**Figure 1**). It was discovered by Pauling and Corey in 1951, the same year as the α helix, and termed 'β' because it was the second structure they elucidated. The basic element is a 5-to 10-residue unit of the polypeptide, whose backbone is almost fully extended, called a 'β strand', with rotation angle values of about − 120° for Φ and 140° for ψ. It can be considered in some way as a particular helix with only two residues per turn and a translation of 3.4 Å per residue.

A β strand is not a stable structure and it therefore tends to interact with other β strands that either belong to other regions of the same polypeptide chain, distant in the primary structure (intramolecular), or are present in different polypeptide chains (intermolecular). The adjacent β strands, running either with the same N- to C-terminus directions or opposite N- to C-terminus directions, are stabilized by hydrogen bonds formed between the carbonyl groups of one β strand and the amino groups of another β strand, and vice versa.

The corresponding structures contain alternate C^{α} atoms lying a little above and a little below the plane of the sheet and, for that reason, are called β pleated sheets. These pleated sheets are designated either parallel or antiparallel (**Figure 1**) depending on the relative direction of the constituent β strands (2 to 6 strands, on average), or else mixed β sheets when β strands combine with some β strand pairs parallel and some antiparallel. In almost all known protein structures, these different β pleated sheets have a right-handed twist, with more positive values of both the Φ and ψ angles.

As in the case of α helices, some amino acid residues have a higher tendency than others to form β sheets. Thus Val, Ile, Thr, which contain branched side-chains, and the three aromatic residues Phe, Tyr, Trp favour the formation of β sheets. In contrast, Glu, Gln, Lys, Asp, Asn, Cys and Pro exhibit low propensity for forming β sheets.

Aggregation of structure elements

In addition to the repetitive structures – α helices and β sheets – a number of nonrepetitive well-ordered structures are present in protein chains to aggregate the secondary structure elements. They occur mostly at the surface of proteins and frequently contain Gly and Pro because of the special conformational properties of these two residues. They are mainly turns, connections and loops that allow formation of structural motifs that constitute the so-called 'supersecondary structure'. A few examples of these motifs are the helix–loop–helix motif, consisting of two α helices joined by a loop region; the hairpin β motif, composed of two adjacent antiparallel β strands connected by a loop; the Greek key motif containing four adjacent antiparallel β strands; and the beta–alpha–beta motif formed by a β strand followed by a loop, an α helix, another loop and a second β strand.

Prediction of secondary structure

One of the decisive factors that determine the secondary structure of a protein (and higher levels of structure) is its amino acid sequence. The different parameters that influence a protein's stability and structural organization include the size, shape, charge, hydrogen bonding, hydrophobicity and degree of freedom of the side-chains of its amino acid residues. Prediction of the secondary structure from the primary structure generally takes into account not only the individual properties of each residue along the chain but also the overall properties of several adjacent residues. An element of secondary structure is more likely to form when several contiguous residues favour that structure.

Several predictive methods have been proposed, among which the method of Chou and Fasman, based on statistical analysis of an X-ray structure database, is the simplest and the most widely used. However, accurate prediction of protein structure from sequence is still an unsolved problem.

Tertiary Structure

The tertiary level of structure refers to the spatial arrangement of a polypeptide chain through folding and coiling to produce a compact globular shape. For some proteins the participation of molecular chaperones is required. The tertiary structure is essentially determined by the packing of the secondary structures, α helices and β sheets, which combine to form one or several units called 'domains'. These combinations are limited in number, and some of them are especially frequent in proteins. They represent the fundamental elements of globular polypeptide chains in terms of three-dimensional structure as well as in terms of function. On average, a single domain consists of 100–150 amino acid residues, corresponding to a globule of about 25 Å diameter. Some domains can be isolated as fragments by limited proteolytic cleavage of the linking peptide chain. Such fragments keep the original conformation they have in the native protein; they are stable and can fold/refold like autonomous structures.

Classification of protein structures

According to the types of structural motifs that constitute the domains, protein structures can be classified into four main groups:

1. Proteins, built up exclusively from α helices in an antiparallel or perpendicular arrangement (e.g. myoglobin, haemoglobin, ferritin, bacteriorhodopsin, phospholipase C).
2. Proteins, consisting exclusively of β sheets in which the β strands are arranged predominantly in an antiparallel fashion (e.g. chymotrypsin, rubredoxin, immunoglobulin, superoxide dismutase, concanavalin A).
3. α/β Proteins, containing β sheets, mostly parallel, surrounded by α helices (e.g. alcohol dehydrogenase, subtilisin, thioredoxin, hexokinase, carboxypeptidase A).
4. α + β Proteins, constructed from a combination of α and β motifs distant from one another in the amino acid sequence (e.g. thermolysin, insulin, papain, lysozyme, ribonuclease A).

The spatial structure of proteins can be determined using a variety of techniques. Thus, spectroscopic techniques (ultraviolet, infrared, fluorescence) provide information on the location and interactions of certain chemical groups. Circular dichroism spectroscopy can be used to analyse the secondary structural components. To obtain detailed data on the arrangement of atoms within a protein, the two main techniques are X-ray crystallography and two-dimensional nuclear magnetic resonance. Thus, the first determination of the complete three-dimensional structure of a globular protein, the sperm whale myoglobin, was achieved in 1958 by John Kendrew on the basis of X-ray crystallography at low resolution.

Fibrous proteins versus globular proteins

An intermediate situation between pure secondary structures and the tertiary structures of globular proteins is represented by the class of fibrous proteins. These proteins have a regular and extended structure that is due to repeated amino acid sequences. They are composed of individual polypeptide chains that are often laterally crosslinked and thus yield physically resistant and water-insoluble structures.

For instance, silk fibroin, a protein of the β-keratin type, is a high-molecular weight protein consisting of a series of antiparallel β-sheets interspersed with irregular peptidic segments and packed together by loose linkages, which renders the silk pliant. Several proteins that play important structural roles in cells consist of two or three right-handed α helices wrapped around one another in a left-handed coil. Examples are α-keratin (a cysteine-rich protein of hair, nails, feathers and fur), myosin, and fibrinogen.

In collagens (proteins of skin, cartilages, bones and blood vessels), no α helix can form because of their high content of glycine, proline and hydroxyproline, which appear in repetitive sequences. Each polypeptide chain has a twisted threefold helical conformation that is left-handed. Three such polypeptide chains are coiled together and stabilized by hydrogen bonds to give a triple helix. On heating, this triple helix unfolds and generates gelatin.

α Helices may be converted into β sheets, as for example occurs when the α-keratin of hair or wool is stretched in a moist and hot atmosphere. The conversion of α helices into β sheets may also explain the abnormal folding of some proteins that tend to aggregate in certain disorders such as Creutzfeldt–Jakob (prion), Alzheimer (β-amyloid) and Parkinson (α-synuclein) neurodegenerative diseases.

Quaternary Structure

Many proteins are made up from two or more polypeptide chains, called subunits or monomers, which may have identical or different amino acid sequences. Such polypeptide aggregation, which represents the quaternary structure, is generally of critical importance to the proper functioning of these oligomeric proteins. Indeed, except in a few cases (e.g. aspartate transcarbamylase), no protein activity is observed when the constituent subunits are separated. Each subunit is usually folded independently, then interacts with the other subunits because they display complementary surfaces as far as shapes and physical interactions are concerned.

The tightness of binding is very variable: some aggregates are very stable and hard to dissociate, whereas others are rather labile. The interactions between two identical subunits (a homodimer) are either isologous or heterologous. Isologous association utilizes the same surfaces on both subunits, whereas heterologous interaction involves two different sites.

As in individual subunits, nonpolar interactions occur preferentially at the centre of the interfaces and hydrophilic groups are located at the periphery for interacting with polar solvents. On aggregation, the accessible surface area of each subunit is reduced by about 10–20%. To minimize the free energy in the aggregated form, the subunits are usually packed in a symmetrical fashion, as in crystals.

Summary

Proteins are biological macromolecules of major importance, both quantitatively and qualitatively, in all living organisms. They are constituted from basic units, called amino acids, which are covalently linked together to form the primary structure. The amino acid sequences determine

the higher structural levels of proteins (secondary, tertiary and quaternary) and specify their biological properties.

Further Reading

Branden C and Tooze J (1991) *Introduction to Protein Structure*. New York: Garland Publishing.

Creighton TE (1993) *Proteins*, 2nd edn. New York: WH Freeman.

Hamaguchi K (1992) *The Protein Molecule*. Tokyo: Japan Scientific Societies Press.

Hecht SM (1998) *Bioorganic Chemistry: Peptides and Proteins*. New York: Oxford University Press.

Higgins SJ and Hames BD (1999) *Post-translational Processing*. New York: Oxford University Press.