BIOMOLECULES: INTRODUCTION, STRUCTURE AND FUNCTION

Proteins

M. Y. Khan and Sangeeta Saxena
Department of Biotechnology
B.B. Ambedkar University
Raibareli Road
Lucknow

(October 2006)

CONTENTS

Keywords

Peptide bond, oligopeptide, polypeptide, amino acid sequence, amino acid compostion, primary structure, secondary structure, tertiary structure, quaternary structure, denaturation, renaturation, native state, quasi-native state, conjugated proteins, globular proteins, fibrous proteins, lipoproteins, prosthetic groups, *cis-trans* configuration, Ramachandran plot, protein folding, α-helix, β-pleated sheet, β-turn, mad cow disease, prion, oligomeric proteins, salt-linkages, hydrogen bonding, van der Waal interaction, hydrophobic interaction, solvation, hydrodynamic properties, salting-out, salting-in, keratins, elastin, haemoglobin, myoglobin, collagen, scurvy, peptidoglycan, nucleoproteins, histones, peptide synthesis, Merrifield solid phase synthesis, chylomicron super secondary structures, haem group, hair curls, random coil.

In the middle of the nineteenth century, the time when new breakthroughs in chemistry and physics were being made virtually everyday, a Dutch scientist Gerardus Mulder described the presence of a substance in living tissues as something which was "without doubt the most important of all substances of the living world, and without it life on our planet would probably not exist". On suggestion of one of his friends, the famous Swedish chemist Berzelius, Mulder named this "most important of all substances" as protein (derived from the Greek word *proteios,* meaning "holding first place" or of the prime importance) in 1838. Proteins are indeed a group of cellular components without which life on the planet earth is unimaginable. We shall seldom come across a class of molecules capable of performing as diverse functions as proteins do.

Proteins are complex molecules made up of simple organic molecules known as amino acids. Amino acids are linearly linked through peptide bonds to form chains of varying lengths. Depending on the number of amino acids present, the chains are named as dipeptide (with two amino acids), tripeptide (with three amino acids), tetrapeptide (with four amino acids) and so on. A small peptide-chain having four to eight amino acids is known as oligopeptide whereas those having between nine to about forty amino acids are simply named as peptides. The name polypeptide is generally used for chain lengths of over forty amino acids. As we shall see later in this chapter, very high levels of complexities in the structure of polypeptide chains are achieved when they bend and fold in three- dimension to acquire unique identities associated with characteristic function(s). Once folded and equipped with some biological function, the polypeptides are bestowed with a new name, proteins.

Classification of Proteins

Since proteins have different amino acid composition (the type and number of each amino acid present in the protein), sequence (the order in which amino acids are linked), size, conformation (three dimensional folding pattern) and function, their classification becomes a rather difficult task. Although all schemes have some overlapping features, depending on their physicochemical properties and functions proteins can be classified in different ways.

Classification Based on Solubility

Proteins differ in their aqueous solubility from being highly soluble to totally insoluble in water. Water-soluble proteins are generally smaller in size and contain proportionally higher amount of polar amino acids. They usually have compact and globular structures and can be coagulated by exposure to heat. Some common examples of water soluble proteins include plasma albumin and the well known egg proteins, namely ovomucoid and ovalbumin. Water-insoluble proteins are usually larger is size, fibrous in nature and contain relatively larger amount of non-polar amino acids. They are more common among plants and are soluble in alcohol and acidic or alkaline solvents. Gliadin from wheat and zein of maize are well known examples of this class of proteins. In addition, there are many proteins that share the properties of both of these classes to varying extent. They are sparingly soluble in water and have variable sizes and diverse functions. Collagen and fibrinogen are well known examples of this class of proteins.

Classification Based on Shape

Proteins are known to have variety of shapes and sizes. Depending on the kind and level of secondary and tertiary structures present in them, they can be classified as globular proteins

having very compact and extensively folded (axial ratio i.e. length to breath ratio <10) conformation as seen in haemoglobins and cytochromes or as fibrous proteins characterized by elongated, filamentous appearance (axial ratio >10) with regular repeats of secondary structures. Keratins of skin, hair and nail; silk protein and collagens are the best examples of this class of proteins. Very often we come across proteins that can not be placed in either of these categories. They have large size and possess irregular structures with limited contents of conventional structures seen in globular proteins. Tenacin and fibronectin are good examples of such proteins.

Classification Based on Composition

Composition of proteins is the most common criteria for their classification. Proteins devoid of structural constituents other than amino acids are called simple proteins. If a protein is associated with a non-amino acid component it is named as conjugated protein. These two groups have been further sub-grouped as follows:

Simple Proteins

Albumins: These are soluble in water and have small to medium size (15-70 kDa) with well defined three dimensional structures. A common example is human serum albumin.

Globulins: They are either insoluble or sparingly soluble in water but soluble in dilute salt solutions. They have medium to large size (45-820 kDa) and posses globular structures. Familiar examples include different types of immunoglobulins.

Glutenins: They are insoluble in water but soluble in alkalies and are generally present as storage proteins in plants. They are rich in gutamic acid content and are heat coagulable. Well known examples include glutanin and oryzenin from wheat and rice respectively.

Gliadins: They are water insoluble but are soluble in alcohol. They are rich in amino acid proline but poor in lysine contents. Hordein of barley, gliadin of wheat and zein of maize are examples of this sub-group of proteins.

Protamines: These proteins have small size, are water soluble and noncoagulable by heat. They lack tryptophan, cysteine and tyrosine. They are highly basic and have strong tendency to combine with nucleic acids. Salmine from salmone and sturine from sturgeon are well known examples.

Scleroproteins/Albuminoids: These are highly stable, low solubility fibrous proteins used in the development of biological structures. Well known examples include α-keratin, elastin and collagen.

Conjugated Proteins

Glycoproteins: Proteins containing covalently linked carbohydrates are known as glycoproteins. Carbohydrate moieties are generally linked with the hydroxyl-oxygen of threonine/serine/tyrosine amino acid residues of the proteins or through the nitrogen of their

histidine/lysine/arginine residues. Well known examples include glycophorin of the membrane and collagens.

Lipoproteins: They contain lipids (triacyl glycerol, phospholipids and cholesterol) as their structural constituents. The amino acid composition of these proteins is dominated by non-polar amino acids. Good examples for this sub-group of proteins include chylomicron and β-lipovitellin of the egg yolk.

Nucleoproteins: These are arginine and lysine rich highly basic proteins, which are non-covalently linked to chromosomal DNA to impart it structural stability. Histones are the best representatives of these proteins.

Metalloproteins: These proteins are characterized by presence of one or more metal ions in their functional structures. Haemoglobin and carboxypeptidase A contain iron and zinc ions respectively as part of their functional structures.

Flavoproteins: Flavin mononucleotide (FMN) and flavine dinucleotide (FAD) are two most common prosthetic groups derived from riboflavin (vitamin B_2). Very firm association of FMN and FAD with certain proteins, generally enzymes, results in the formation of flavoproteins. Glycolic acid oxidase and lipoyl dehydrogenase are examples of FMN and FAD flavoproteins respectively. In cases (for example succinate dehydrogenase) where a functional protein structure is associated with metal as well as FMN/FAD prosthetic groups at the same time, the protein is called metallo-flavoprotein.

Phosphoproteins: These proteins contain phosphoric acid as organic phosphates and have high nutritional value. Well known examples of this class of proteins include casein from milk and ovovitellin from egg-white.

Classification Based on Functions

Life and living activities are very complex and proteins are involved in virtually all forms of its manifestations. The classification of proteins based on their functions is therefore as difficult as the classification of "living activities" itself. However, in terms of the common nomenclature that we generally use in our daily life (like stores, defense, regulator etc.), the proteins can be classified in several groups as listed in Table 1.

Peptides

As mentioned in the beginning of this chapter, two amino acids can be covalently linked together by an amide bond between the α-carboxyl group of one amino acid and the α-amino group of another. This amide bond is called as peptide bond and the resulting product is known as a peptide. Formation of a dipeptide involving alanine and glycine amino acids as its structural constituents, which are referred to as amino acid residues, is shown in Box 1. The free carboxyl group of the above dipeptide (alanylglycine) can react with the amino group of a third amino acid, serine for example, resulting into a tripeptide (alanylglycylserine) containing three amino acid residues. Thus successive reactions, eliminating a water molecule each, would lead to the formation of larger peptides and eventually a protein. The final product will have a free amino-

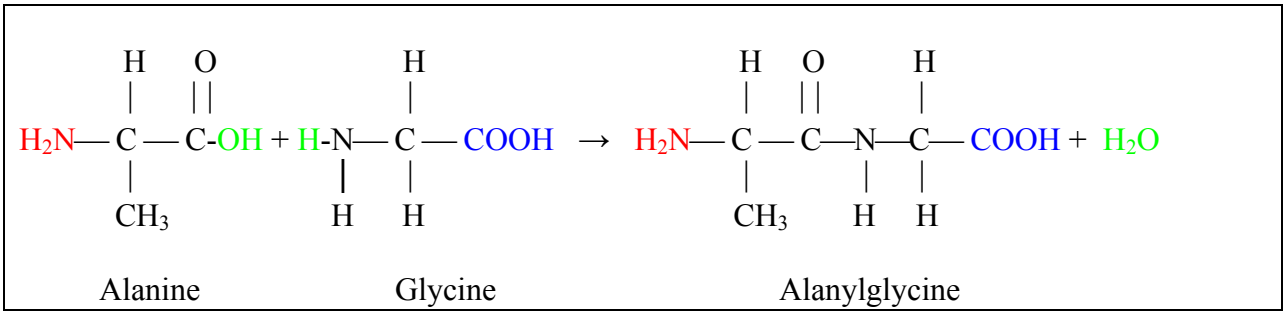
and a free-carboxyl group at the two ends that are named as the amino and carboxyl termini of the peptide/proteins respectively. The amino acid sequence of peptides is expressed in terms of three-letter or one-letter symbols of their constituent amino acids. By convention, the amino acid having free amino group represents the amino terminus and is considered to be the first amino acid of the chain while the one that represents the carboxyl terminus is considered to be the last. Hence the sequence of alanylglycylserine could be written as: Ala-Gly-Ser or AGS.

Table 1: Functions of proteins

Protein Class	Functions	Examples
Structural Proteins	They are used as bricks and mortars to construct the biological buildings and machineries	α-Keratin of fur, feathers, hairs and claws; collagens of skin, bone and cartilage
Carrier Proteins	They carry metabolites from one site to the other to make biological processes a reality	Haemoglobin carries oxygen; transferrin transports iron
Storage Proteins/ Nutrient Proteins	They serve as biological store houses to preserve nutritional proteins which act as source of essential amino acids	Casein of milk, ovalbumin and ovomucoid of egg, and glutelin of wheat, ferritin
Enzyme Proteins	They act as biological catalysts and make an otherwise slow or improbable reaction fast and feasible	Digestive enzymes trypsin and pepsin, papain from papaya and ribonucleases
Hormone Proteins	They act as biological signals; mediate and regulate physiological processes	Insulin, glucagons and adrenocorticotrophic hormone
Defense Proteins	Protect against foreign invaders like bacteria and viruses, make survival possible under hostile conditions	Antibodies, thrombin, antifreeze proteins and lysozyme in tears
Proteins as Toxins/ Poisons	They are toxic/poisonous to others but provide a defense tool to organisms they belong to	Snake venoms, diphtheria toxin, ricin in caster bean, gossypin of cotton seed

Physicochemical properties of peptides depend both on their amino acid composition as well as the amino acid sequence. Chemical properties of the constituent amino acid residues are distinctly manifested in their peptides. There are many naturally occurring peptides that play crucial physiological roles. Glutathione, a tripeptide (glutamylcysteinylglycine), for instance, provides reducing environment in critical biological reactions. Similarly, oxytocin is made up of eight amino acid residues and acts as a hormone that helps in milk ejection and also affects

uterine muscle before the parturition period. Other physiologically important peptides include angiotensins, gastrin, bradykinin, β-corticotropin and carnosine.



Box 1: Formation of a peptide bond

The Structure of the Peptide Bond

The peptide bond is the covalent bond that links amino acids together. It is essentially an amide linkage having certain characteristics that are very important to protein structure. The values of bond lengths and bond angles of various linkages related to a peptide bond are shown in Figure 1. It is evident that the length of the C—N bond (1.32 nm) is not typical of a true single bond (1.47 nm) seen in other compounds. This is explained by attributing a substantial double-bond character to the C—N bond acquired by resonance hybridization as shown in Box 2.This delocalization of π electrons as shown in Figure 2 imparts a partial double bond character to the C—N bond restricting free rotation across it. This results into an arrangement wherein all six atoms about the peptide bond lie in the same plane as shown Figure 1. Although the group of atoms about the peptide bond lie in a relatively rigid plane, they still have the option to exist in two possible configurations namely, *cis* and *trans* (Box 3). Generally, the peptide bonds in proteins are formed using naturally occurring twenty L-amino acids and are present in *trans* configuration because it is more stable arrangement, that avoids sterical interference among the bulky groups.

Protein Structure

As mentioned earlier, the peptide bond acquires planer geometry where free rotation across the C—N bond is restricted. Owing to their true single bond nature, rotations about the covalent bonds that connect each C_α to the adjacent planer peptide groups is permissible. Such rotations about the C_α–N and C_α–C bonds are measured by Φ (*phai*) and ψ (*psi*) angles respectively (Figure 3). In principle, Φ and ψ angles can have any value between -180^0 and $+180^0$ and therefore, all possible conformations of the polypeptide chain can be described in terms of their Φ and ψ angles. In other words, every possible set of one value each of Φ and ψ angles in the -180^0 to $+180^0$ range will signify one conformation of the protein.

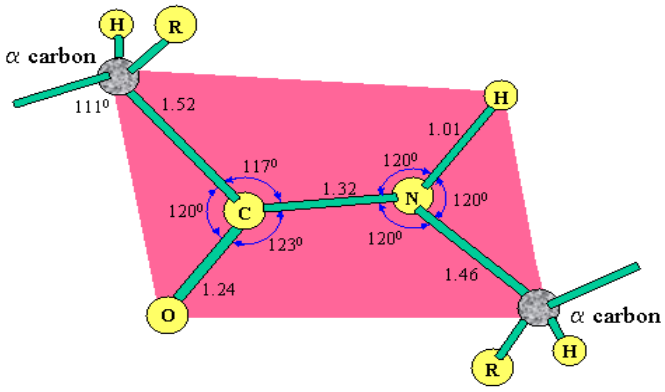
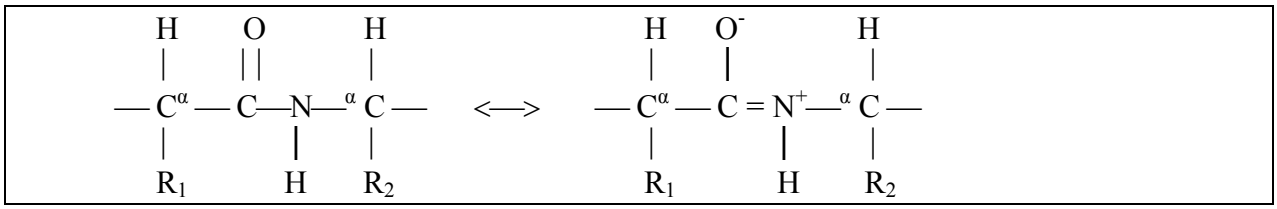


Figure 1: A peptide bond defining bond angles and bond lengths.



Box 2: Resonance stabilization of peptide bond

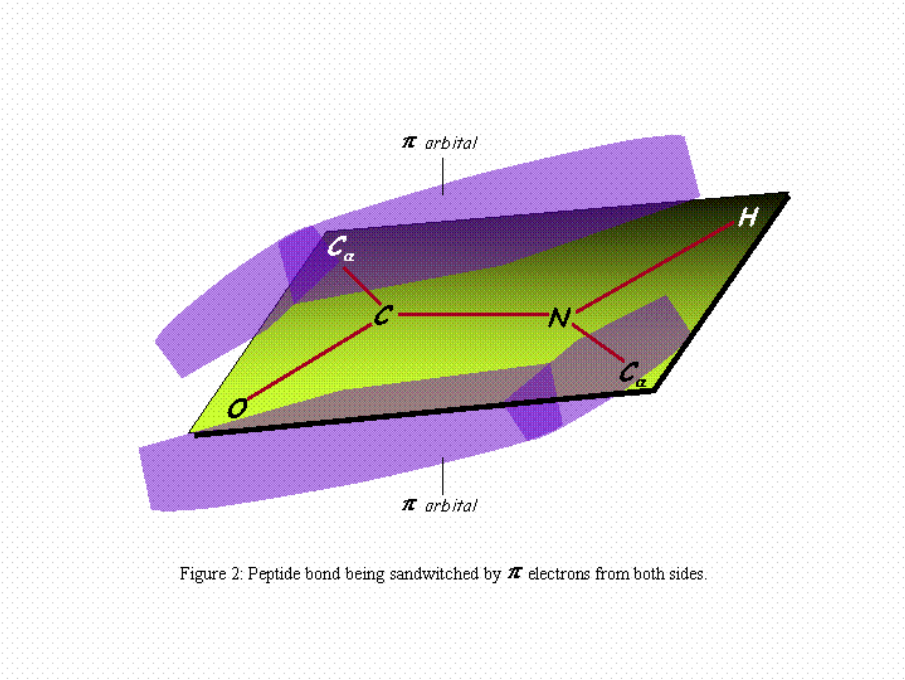
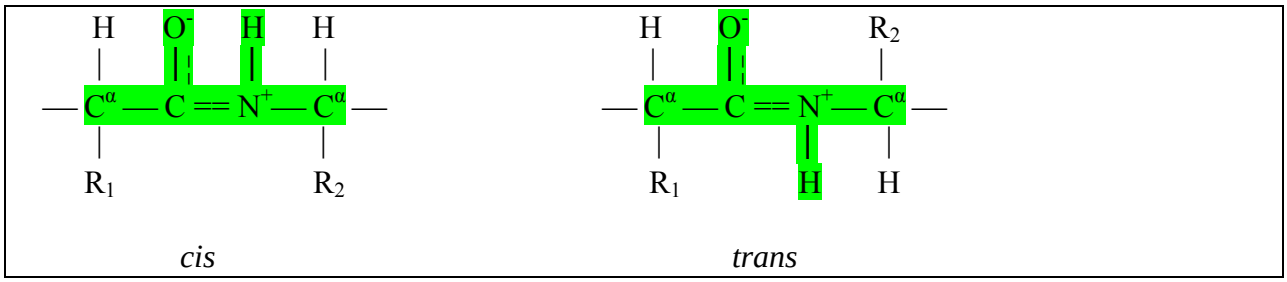


Figure 2: Peptide bond being sandwitched by π electrons from both sides.



Box 3: The *cis* and *trans* configurations of peptide bond

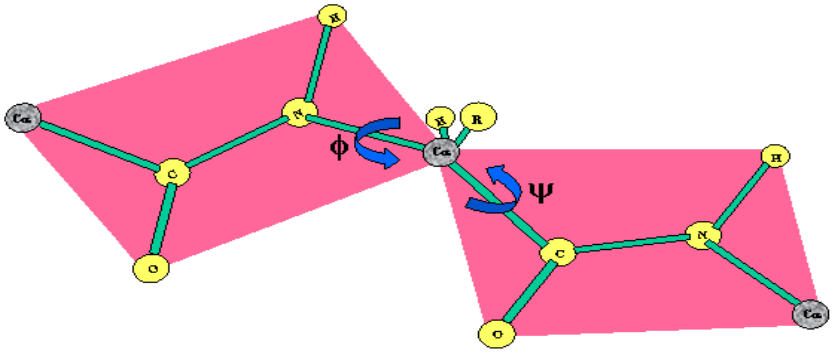


Figure 3: ψ and ϕ angles are used to express rotations across the Cα–C and Cα–N bonds respectively.

Ramachandran Plots

Before proceeding further, let us first examine the conventions used to name the direction of rotations about the Φ and ψ angles. The reference point is always considered an imaginary conformation where Φ and ψ angles are taken to be zero and both the peptide planes connected to a common C_α atom lie in the same plane. On looking from the C_α in either direction, clockwise rotations are considered positive and accordingly the Φ and ψ dihedral angles will have a positive signs. Similarly, an anti-clockwise rotation would assign negative values to the two angles. It is thus evident that if both the dihedral (Φ and ψ) angles are assigned a value of + 180^0 each, the polypeptide chain would acquire a fully extended conformation. If , on the other hand, the two angles are assigned a value of 0^0 each, the two successive peptide planes would become co-planer and come dangerously closer to each other to the extent that the carbonyl oxygen and amino hydrogen would sterically clash. It is therefore conceivable that not all possible structural conformations will ever become a reality. This was for the first time realized and mathematically analyzed by the Indian scientist, G.N. Ramachandran. Since each polypeptide structure can be fully defined by its characteristic pair of Φ and ψ angles, the

backbone conformation of any particular residue in a protein can be represented as a point on a plot of Φ *versus* ψ angles. Such plots are called Ramachandran plots because G. N. Ramachandran was the first to make extensive use of these plots to analyze the structure of proteins. Some features of a typical Ramachandran plots are shown in Figure 4. Clearly, well known secondary structures (described below) of proteins have a tendency of acquiring only certain "allowed" values of the two dihedral angles. A major area of the plot is represented by such combinations of Φ and ψ angles which are "disallowed" while limited regions may also be partially allowed conformational zones.

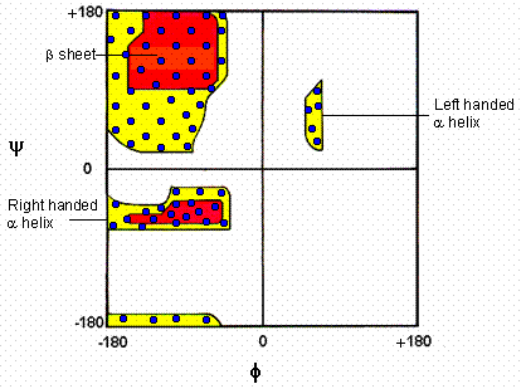


Figure 4: Ramachandran plot of ϕ and ψ angles showing allowed (shown by dotted areas) and nonallowed combinations (undotted areas) of these two dihedral angles.The figure has been downloaded from source http://www.cryst.bbk.ac.uk/pp and adopted after suitable modification.

Levels of Structure in Protein Architecture

Using the feasibility of unrestricted rotations across the aforesaid two bonds as also their covalent back bones, proteins acquire their characteristic conformations. While reaching their destined natural conformation, a process known as protein folding, proteins follow all physicochemical laws faithfully. The protein structure can be said to be of four different types (Figure 5). Although there are situations when the description of one type of structural level overlaps with that of the other, the scheme for the nomenclature of different structural levels still works satisfactorily.

Primary Structure

All proteins have well defined orders in which their amino acids are linked to generate the chain like structure. This characteristic sequence of amino acids in a protein is called its primary structure. This is the most fundamental level of protein structure. A protein may contain one or more than one polypeptide chains having similar or different primary structures. If present, these chains are associated with each other with the help of weak secondary interactions or disulfide bonds (see determination of the amino acid sequence of a polypeptide chain). Higher levels of structural organizations in protein architecture depend on its primary structure. Even a small change in the primary structure of a protein has a potential of changing its function substantially.

The example of sickle cell hemoglobin illustrates this fact beautifully. The primary structures of normal adult and sickle cell haemoglobins are the same at all but one position. Single substitution of a glutamic acid (Glu) residue present in the normal haemoglobin with that of a valine (Val) amino acid residue alters the property of the protein to the extent that the longevity of the person having sickle-cell disease is significantly affected (Box 4).

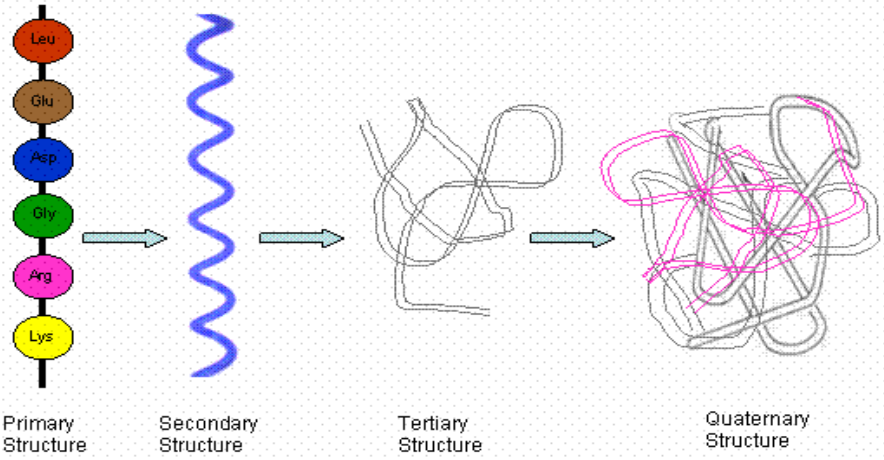


Figure 5: Different levels of protein structure.

Normal	NH_2–Val–His–Leu–Thr–Pro–Glu–Glu–Lys…….
Sickle Cell	NH_2–Val–His–Leu–Thr–Pro–Val–Glu–Lys…….

Box 4: Partial primary structure of β-chain of normal and sickle cell hemoglobin

Secondary Structure of Proteins

All atoms of a protein polypeptide chain have specific positions in its three dimensional structure. This three-dimensional structure is decided by the above referred Φ and ψ angles after taking into account different steric constraints as also the intra and inter chain atomic interactions. These weak interactions, to be discussed later, include hydrogen bonds, electrostatic, hydrophobic and van der Waal interactions. The three dimensional folding of proteins clearly has two levels of structural hierarchy. First, the polypeptide chain locally twists

in stretches of helical or sheet like structures resulting into regular folding patterns known as secondary structure. Secondly, the locally folded structures of the entire polypeptide chain reassemble in three-dimension to generate what we call the tertiary structure of the protein.

The different kinds of secondary structures can be considered as the various stages of a spring being stretched progressively. Each stage of the stretching spring having a characteristic diameter and linear length is analogous to different types of secondary structures. The formation and subsequent stabilization of some common secondary structures are described below.

Helices: All coiled/helical structures will have characteristic number of amino acid residue in its one full turn, *n*; pitch (i.e. the linear distance required by the helix to repeat itself or the linear distance covered by one turn), *p* (Figure 6); magnitude of rise or the increase in the helix linear-length per amino acid residue, *h*. Working with molecular models in 1950s, Pauling and his research collaborators noticed that proteins prefer to acquire a small number of regular conformations. Among these conformations, a right handed coiled structure, named α-helix, was particularly favourable for a stable structure because it had minimum possible steric hindrances with very high potential for hydrogen bond formation. It has an average of 3.6 amino acid residues per turn of the helix and is stabilized due to hydrogen bonding between the –C=O group of an amino acid and the –N-H group of the fourth amino acid ahead in the chain. The values of the Φ and ψ dihedral angles of α-helix are -57^0 and -47^0 respectively. The linear distance covered by one full turn of the helix is 5.4 Å that translates to a rise/amino acid residue of 1.5 Å (5.4÷3.6 = 1.5).

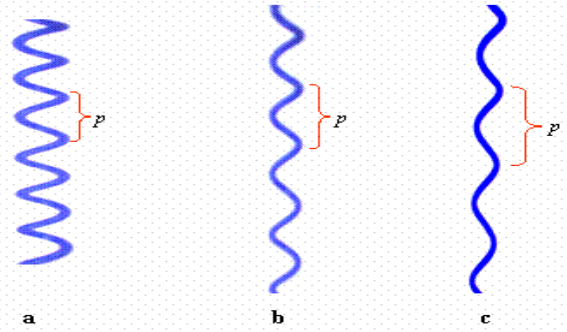


Figure 6: Helical structures (a,b,c) found in proteins. It may be noted that the number of amino acid residues in its one full term and other characteristics, for example *p* (the distance covered by one turn of the helix) vary according to the type of the helix. *p* is 5.6 Å for the α helix.

Although α-helix is the most stable and prevalent helical structure found in proteins, other helical arrangements like л and 3_{10} helices has also been detected. Like α-helix, these helices are also stabilized by intrachain hydrogen bondings. Each carbonyl oxygen is hydrogen bonded to the imino hydrogen of the third residue up in 3_{10} helix. The 3_{10} helix is not as common as the α-helix

and it has *n*, *p* and *h* values of 3.0, 6.0 Å and 2.0 Å respectively. Although sterically possible, the л helix has not been observed in proteins. The values of the helix-defining parameters, *n*, *p* and *h*, for л helix are 4.4, 5.28 Å and 1.2 Å respectively.

Pleated sheets: A helical structure having fewer than three amino acid residues per turn can not be sustained because the linear hydrogen bonds between residues in the same chain can not be easily formed. The "helix" then becomes overstretched and acquires an extended sheet like structure known as β-pleated sheet as shown in Figure 7. Every successive amino acid residue rotates by 180^0 with respect to the preceding one in such a manner that each "pleat" accommodates 2 residues. The structure is stabilized either by inter-chain hydrogen bonding between two different but adjacent polypeptide chains or by intra-chain hydrogen bonds between two stretches of the same polypeptide chains brought closer to each other by its bending. As has been noted on page 6, the polypeptide chains have directionality in the sense that they start with NH_2-termini (or simply N-termini) and end with the COOH-termini or the C-termini. Based on this convention, β-pleated sheets have been classified in to two types. If the two chains linked with hydrogen bonding have the same N→C directions, the sheet is referred to as parallel β-pleated sheet. If however, the two polypeptides run in the opposite directions, the structure is called anti-parallel β-pleated sheet (Figure 7).

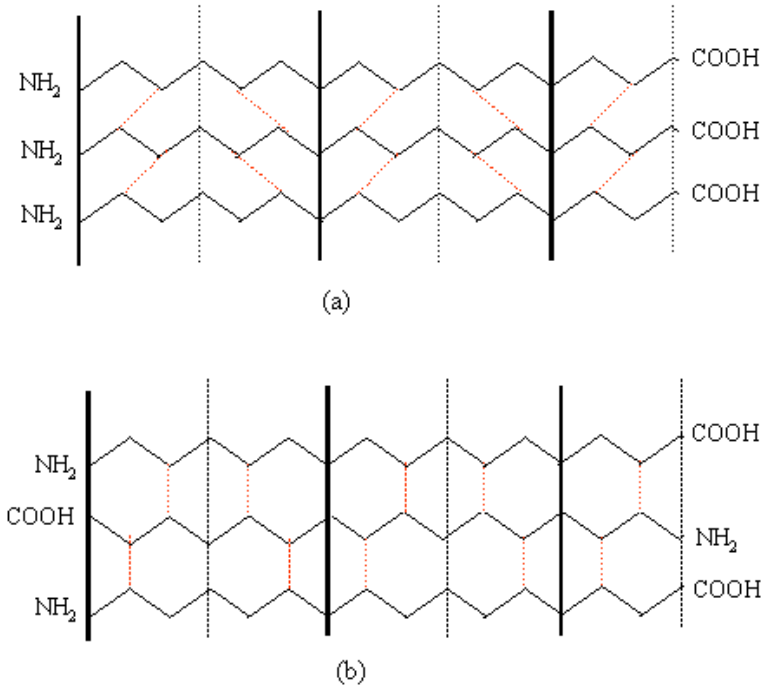


Figure 7: Parallel (a) and antiparallel (b) β-sheets in proteins. Thick and thin lines have been put in to indicate the regions of the sheet nearer and farther from the viewer respectively.

β-Turns: Being an amino acid with the smallest side chain (hydrogen), glycine has the potential of getting accommodated in structures normally not allowed for other amino acids as predicted by Ramachandran plot described above. Thus glycine can be easily fitted in bends that is generally not possible with other amino acid residues. These sharp turns are called β-Turns. The structure of proline residues is also unusual as it contains a ring-structured side chain that prevents rotation about C_α-N bond imparting a fixed value to the Φ angle at about -65^0. This ring structure can be accommodated in β-turns, which have also been found to have preference for serine residues. β-Turns thus represent very short stretches of four amino acid residues that provide reversal in the direction of the polypeptide chains and therefore play an important role in their packing. The structure is stabilized by hydrogen bonding between carbonyl group of an i^{th} amino acid residue and the amide hydrogen of the residue $i+3$ ahead in the chain. Such bends and turns often occur at the surface of proteins.

Super secondary structures: Structural entities originating from different combinations of the above described conventional secondary structures are called super secondary structures. These structures are characteristically present in globular proteins and are considered to be of three types namely, β-α-β type (two parallel β-pleated sheets connected by an intervening α-helix), Greek key type (antiparallel β-pleated sheets connected by random structures to give the appearance of the classical Greek pottery design) and β-Meander (five antoparallel β-pleated sheets connected by β-turns).

Tertiary Structure of Proteins

We have seen so far that how proteins acquire various types of secondary structures based primarily on the special characteristics of their peptide linkages. The secondary structure is relatively simple with fewer motifs. The functional three- dimensional structures of proteins are far more complex and require "assistance" from multiple sources for their genesis. The diversity in the nature of R groups (the side chain) of the twenty amino acids plays an important role in the self-assembly of proteins. These side chains interact among themselves on one hand and with the water and other components of the cellular medium on the other. All these interactions and also the factors that were instrumental in the formation of secondary structures, assist the protein polypeptide chain in undergoing extensive coiling and folding to generate a relatively compact and globular conformation known as the tertiary structure of the protein (Figure 5). Tertiary structure is the unique characteristic of every protein and is generally associated with a distinct biological function. The information required for the genesis of tertiary structure of a protein is inherent in its primary structure. Amino acid residues that are well separated in the primary structure often come close to each other in the tertiary structure.

While folding to a compact and globular conformation, proteins strictly follow the laws of thermodynamics. Although the folded structure is only marginally stable over the corresponding unfolded state, the delicately balanced cellular machinery always ensures that the newly synthesized polypeptide chain folds to its predestined functional three-dimensional structure *in vivo*. Any failure on the part of the nascent polypeptide chain to fold to the correct conformation may lead to serious clinical problems. One of the best such known example is bovine spongiformform encephalopathy or "mad cow disease". A protein known as prion related protein (PrP) is present in many animals, including human beings, in a non-pathological form called prion-related protein cellular (PrP^c). Under certain circumstances, PrP^c changes its tertiary

structure to a different form designated as PrP^{sc} (prion-related protein scrapie). PrP^{sc} is pathogenic and badly affects the nervous system. On ingestion, PrP^{sc} can induce conversion of PrP^{c} to PrP^{sc} in the recipient. This makes mad cow disease infectious. This discovery enabled its discoverer Stanley Prusiner won the Nobel Prize in 1997.

Quaternary Structure of Proteins

Many proteins have more than one polypeptide chains that are assembled together to form their functional structure. Such proteins are called oligomeric proteins. The number of these polypeptide subunits present in a protein and their relative arrangement in the three-dimensional structure is referred to as quaternary structure. Insulin and haemoglobin, having two and four polypeptide subunits respectively, are well known examples of oligomeric proteins having quaternary structure. Depending on the nature and the structural and functional complexity, the quaternary structure may comprise several subunits as seen in the case of tobacco mosaic virus. Presence of quaternary structure enables proteins to acquire special functional properties. Allosteric property of haemoglobin, for instance, is attributed to its four-subunit quaternary structure. Similarly, the regulation of the activity of lactose synthase, the enzyme that synthesizes the milk sugar lactose, has been shown to depend on its quaternary structure.

The tertiary structure of different subunits of an oligomric protein is generally formed before their association to form the quaternary structure. However, the folding of polypeptides in three dimension (tertiary structure) and their assembly to the functional quaternary structure may also proceed at the same time in certain cases. Thus there appears to be a hierarchy in generation of different levels of protein structure. The primary structure, which has the minimum complexity in this hierarchy, decides the fate of secondary structure having higher level of complexity. Different elements of the secondary structure fold together to generate the tertiary structure of polypeptide-subunits that in turn combine to yield the ultimate level of hierarchical complexity seen in quaternary structure. Interactions that help in the formation and sustenance of tertiary and quaternary structures are the same and would be described in the following section.

Forces Stabilizing the Tertiary and Quaternary Structures of Proteins

A number of interactions are responsible for the formation of various structural elements in proteins. These interactions are very weak as compared to covalent bonds present in biological compounds. The strength of these interactions, which lies generally in the range 0.1 – 6.0 kcal/mole, compares very poorly with bond energies associated with covalent bonds (Table 2). In spite of their small individual contributions, astronomically large numbers of weak interactions contribute hugely to the overall molecular energy associated with a protein molecule. The stabilization of the native folded conformation of proteins over its unfolded polypeptide structure is, however, only marginal. The overall free energy change on folding is negative and lies in the range –2 to –20 kcal per mole. This negative free energy change is the result of a number of factors that are briefly described below.

Electrostatic Interaction

These interactions can occur between positive and negative charges contributed by the two terminal amino acids of the protein. The side chains of acidic and basic amino acid residues (like the β and γ carboxyl groups of aspartic acid and glutamic acid respectively) and the ε groups of

lysine residues are also involved in such interactions. These interactions are analogous to what we see when an acid and its conjugate base interact and therefore they are also some time referred to as salt linkages. The energy of interaction (E) of such salt linkages depends on the number of positive and negative charges, Q_1 and Q_2, and the distance, R, which separates them in the following manner:

$$E = \frac{Q_1 Q_2}{DR}$$

Where D is the dielectric constant of the medium (predominantly aqueous under cellular conditions). Since the value of D is lower in the case of interacting charges that are buried in the protein interior, the E would have higher values for such interactions when they become operative in folded proteins.

Table 2: Energies of different kinds of covalent bonds and secondary weak interactions relevant to biological systems

Bond/Interaction Type	Bond Energy	Bond/Interaction Type	Bond Energy
C—C single bond	82 kcal/mole	C—O single bond	84 kcal/mole
C=O double bond	164 kcal/mole	C—N single bond	70 kcal/mole
C=N double bond	147 kcal/mole	N—H single bond	94 kcal/mole
C—H single bond	99 kcal/mole	C=C double bond	147 kcal/mole
C=S double bond	108 kcal/mole	S—H single bond	81 kcal/mole
S—S single bond	51 kcal/mole	O=O double bond	96 kcal/mole
O—H single bond	110 kcal/mole	O—O single bond	34 kcal/mole
Hydrogen bond	1-3 kcal/mole	Electrostatic Interaction	2-6 kcal/mole
van der Waal's Interaction	0.1-0.2 kcal/mole	Hydrophobic Interaction	< 6 kcal/mole

Hydrogen Bonding

The hydrogen bonding is the name given to interactions involving a fully defined charge and an induced dipole wherein one of the interacting partners is a hydrogen atom. Alternatively, it could also be an interaction involving two dipoles. The electrons of hydrogen atoms attached to electronegative atoms like oxygen or nitrogen have a tendency to marginally shift towards the bonded partner with higher electronegativity. This tendency induces the formation of a dipole with partial positive and partial negative charges on the hydrogen and the other bonded atom (oxygen/nitrogen) respectively. The induced dipoles thus formed interact with other dipoles or charges exactly the same way as it happens in electrostatic interaction. Further, the strength of hydrogen bonds also depends on the angle of the hydrogen bonding atoms and the atom that induces the dipole; maximum strength is attained when all these atoms are co-linear. The energy of interaction in hydrogen bonding is more sensitive to distance than what we have seen in charge-charge interaction above (Box 5). As expected, the hydrogen bond strength would be depending on the dielectric constant of the medium as described above.

Energy of hydrogen bonds between charged atoms and induced dipoles is proportional to: $1/R^2$
Energy of hydrogen bonds between two induced dipoles is proportional to: $1/R^3$
Energy of attractive van der Waal's interaction is proportional to: $1/R^6$
Energy of repulsive van der Waal's interaction is proportional to: $1/R^{12}$

Box 5: Relation between bond energies and the distance of interacting atoms

van der Waal Interaction

Albeit weakly, the interactions between uncharged molecular groups and atoms can also make significant contributions in the formation of protein structures. van der Waal interactions come into play due to induced dipole moments that arise from the fluctuations in the electron charge densities of neighboring non-bonded atoms. These interactions are extremely weak and have strength of only about 0.1 to 0.2 kcal/mole. They can be both attractive as well as repulsive in nature. The repulsive van der Waal forces are generated when electronic charge clouds between molecules begin to overlap. These interactions are highly sensitive to the distance that separates the interacting atoms (Box 5). Although van der Waal interactions are extremely weak in nature, their universal presence and large numbers makes them significant contributor in stabilization of protein structures.

Hydrophobic Interaction

Of all the interactive forces we have discussed so far, the hydrophobic interaction is the least understood factor. It is an entropy driven interaction and is attributed more to the properties of the medium where it operates. The term hydrophobic, which means the "water hating", is misleading since its genesis depends on the unfriendly behaviour of water against certain molecules. Owing to its polar nature, water is a very well organized stable structure in the sense that it has extensive intermolecular hydrogen bonding. Solubilization of a solute in water is favoured only when it can compensate for the gain in energy (i.e. destabilization) of water resulting due to breaking of some of the intermolecular hydrogen bonds for accommodating the incoming solute. Polar or charged solutes like sodium chloride can easily meet this condition and are therefore soluble in water. Since they cannot form hydrogen bonds with water, non-polar substances like methane cannot compensate for the energy gain on account of breaking of intermolecular hydrogen bonds of water that would accompany the solubilization process. When a situation arises wherein a non-polar substance is placed in water, the latter would intend to "expel" all such substances out of its territory (Figure 8). This repulsion tendency of water molecules compels all non-polar or hydrophobic molecules to come in contact with each other giving an impression as if they are interacting with each other. This phenomenon is known as hydrophobic interaction.

Proteins have many hydrophobic amino acids like leucine, methionine, isoleucine, valine and phenyl alanine. On coming in contact with water, these amino acids adjust the protein conformation to secure a non-water zone for themselves. The protein molecules thus acquire a conformation where all hydrophobic interactions are satisfied to the best extent possible. The magnitude of these forces is smaller as compared to electrostatic and other polar interactions.

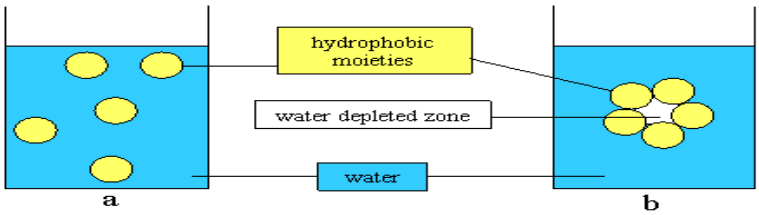


Figure 8: The genesis of hydrophobic interaction. Hydrophobic molecules scattered randomly in water (a) are compelled to acquire an arrangement as shown in (b) where all hydrophobic moieties have minimum contact with water.

Denaturation and Renaturation of Proteins

The term denaturation has been used in different context on different occasions. At the time when not much was known about protein structure, the words denaturation, coagulation and precipitation were changed interchangeably. With increasing knowledge on protein structures and the principles that govern them, the term denaturation is now almost exclusively used to denote structural alterations in proteins. As has been already described earlier in this chapter, proteins generally acquire their characteristic functional properties only after undergoing extensive twisting, bending and folding to a compact three-dimensional structure known as the native state. On treatment with a number of physico-chemical factors like heat, radiations, harsh physical treatments like violent shaking, acid or alkaline solutions, detergents, and chemical compounds like urea and guanidine hydrochloride, proteins generally lose their native structure which is invariably accompanied by loss of the function as well. This phenomenon of loss of the native structure is known as protein denaturation and the factors that cause it are referred to as denaturants. Very often the removal of denaturants enables the proteins to refold back to their functional native state – a process known as renaturation (Figure 9).

Denaturation is generally not associated with loss of covalent structures and is confined to removal of weak secondary interactions instead. The denatured state that is formed as a result of denaturation represents a thread like covalent back bone of the protein polypeptide which is called as random coil. Denaturation always results into loss of biological activity i.e. inactivation of proteins. However, many a time proteins are inactivated as a result of very small or practically no alteration in their structures. This can not be called denaturation. Papain, the protease from papaya, is a good example. When treated with mercury, papain loses all its activity without significant change in its structure. Similarly, there are instances where the compact and globular native state of the protein is inactive by itself and becomes active only after it undergoes a small conformational change upon its binding to certain ligands. Such native forms of proteins are known as quasi-native state. An example to this effect is provided by pyruvate kinase which is activated after its binding to Mg^{++}/K^{+} ions.

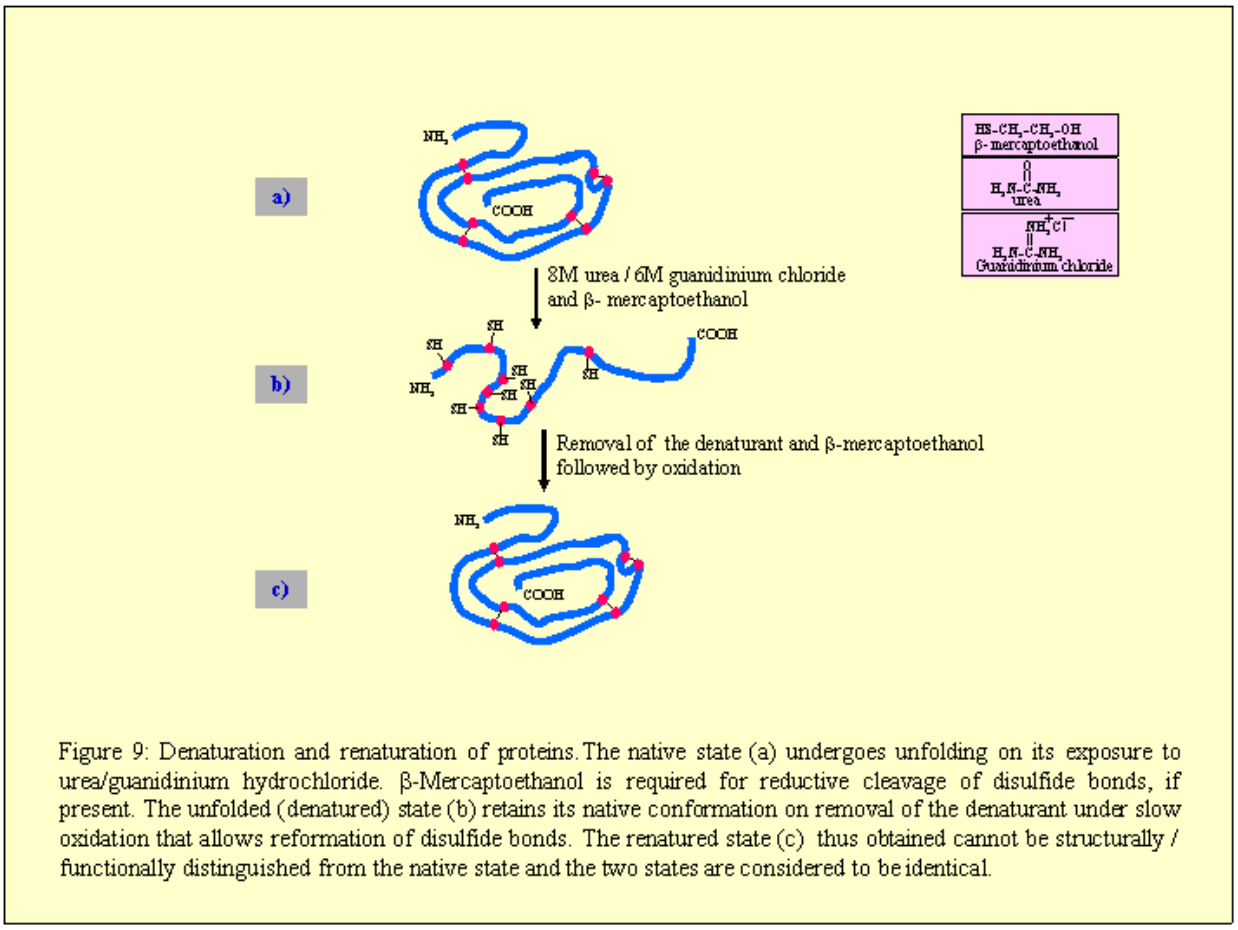


Figure 9: Denaturation and renaturation of proteins. The native state (a) undergoes unfolding on its exposure to urea/guanidinium hydrochloride. β-Mercaptoethanol is required for reductive cleavage of disulfide bonds, if present. The unfolded (denatured) state (b) retains its native conformation on removal of the denaturant under slow oxidation that allows reformation of disulfide bonds. The renatured state (c) thus obtained cannot be structurally / functionally distinguished from the native state and the two states are considered to be identical.

As against coagulation, which is irreversible and involves insolubilization and covalent damage to the structure of the protein molecule, denaturation is generally reversible and does not involve hydrolysis of peptide bonds. Denatured proteins are hard to crystallize and they have lower aqueous solubility than their native counterparts. The digestibility of edible proteins increases upon their denaturation as additional peptide that are buried in the protein interior are exposed and become susceptible to proteolytic attack in the intestine.

Behaviour of Proteins in Solution

Owing to the diverse chemical nature of their structural units i.e. the amino acids, proteins always have some polar groups and are generally charged under physiological conditions. In solution, the polar groups and the charges on the protein molecule provide the basis for their interaction with numerous water molecules. This phenomenon is known as solvation or hydration of proteins. Thus, every protein will have some amount of solvent bound to it resulting in to an increase in its effective molecular size. Similarly, depending on the hydration pattern, the effective shape of the protein in solution would also be different from its shape under un-hydrated condition. The effective volume and shape of proteins under hydrated conditions (i.e. in solution) are known as its hydrodynamic volume and hydrodynamic shape respectively. All physical properties that depend on hydrodynamic shape and volume of proteins are called as hydrodynamic properties. Among others, diffusion, viscosity, sedimentation and frictional coefficients are important hydrodynamic properties that are used for characterization of proteins.

Salting-in and salting-out of proteins

We have already discussed that water has a tendency to favourably interact with every polar/charged species it comes in contact with. In solution therefore, a net attractive electrical potential energy would result when an ion having a given charge is surrounded by ions of the opposite charge. This electrical potential will increase the theoretical molar free energy of the dissolved salt molecules present in ideal conditions where they do not have any interaction. This enhancement in the electrical potential is expected to increase further with increase in ionic strength. The activity coefficient of an ideal solution is always taken to be unity and it would decrease in the presence of added salt under these circumstances. Since the solubility product for a charged-dissolved ion is directly proportional to its activity coefficient, the solubility of proteins increases when smaller amounts of salts are added to its solution and it goes down in presence of excess salt. This increase in the solubility of proteins with added salt is known as salting-in (Figure 10).

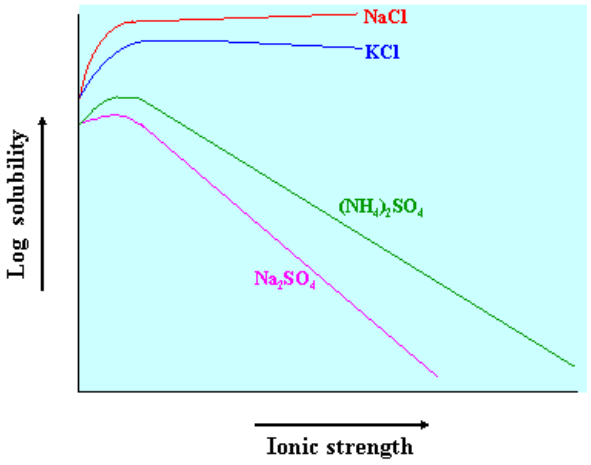


Figure 10: Effect of increasing salt concentrations (expressed in terms of ionic strength) on the solubility (expressed in logarithmic terms) of proteins. Whereas sodium chloride and potassium chloride are more effective in promoting "salting in", the sulfate salts of ammonium and sodium are potent "salting out" agents and are therefore frequently used in salt fractionation of proteins.

In more simple terms, the solubilization can be defined as energetically favourable (negative free energy change) interaction between a solvent and the solute. Factors that increase the magnitude of this interaction enhance the solubility of the solute. Conversely, insolubilization or precipitation refers to a condition where favourable solute-solvent interaction decreases and/or solute-solute interaction increases. When a protein is solubilized in pure water, the latter reluctantly (as noted earlier, water is a well organized structure) diverts some of its intermolecular interactions to water-protein interactions (i.e. solubilization). This water-protein interaction may not prove out to be effective enough to eliminate all protein-protein interactions

(i.e. insolubilization/precipitation or reduced solubility in general terms). This would especially hold good for proteins having relatively fewer polar/charged and more hydrophobic amino acid residues. Addition of small amount of salt in the protein solution would tend to mask some of the protein charges that were promoting protein-protein interaction. This would result into an increase in solubility of the protein or the salting-in. The condition would be totally different when an excess of salt is added to the protein solution. In addition to masking the charges on the protein molecules, the salt-ions having high charge intensity would attract large number of water molecules for their own solvation (note that it would be to the liking of water). The net result is that water molecules are effectively removed from the solution in the sense that they are no more available for solvating the protein molecules. Thus, at very high salt concentration, so much water is engaged with salt ions that the effective protein concentration is increased to the level that it precipitates out of solution. This phenomenon of precipitation of proteins at high salt concentration is known as salting-out of proteins. Since every protein has its own characteristic salt concentration at which it precipitates, this property has been extensively used in purification of various proteins.

Structure and Biological Functions of Some Main Groups of Proteins

Now that we have some understanding of the different kinds of structural motifs present in proteins and also the principles that govern them, we can look for greater details of the correlations that exist between their structure and functions. Few examples from each major class have been taken and an analysis of the data available on their structure and function is presented in the following section.

Fibrous Proteins

These proteins are usually water-insoluble and have elongated rod like appearance. Their primary function is to provide structural framework to the cells and the tissues. Depending on the specific needs, the filamentous shape of these proteins is generated by using one of the standard secondary-structural motifs we have already discussed. Let us now consider a few examples to illustrate some features of these proteins.

Keratins

Keratins are themselves a diverse class of fibrous proteins. They are found in a variety of biological tissues such as hair, horn, scales, beaks, wool, hooves, claws and nails. The two most abundant forms of keratins are called α-keratins and β-keratins. The α-keratin is the main constituent of hair, wool, skin and finger nails. These keratins have an intermediate filament size and their structure is predominated by the α-helical structural motif. Long stretches (over 300 amino acid residues) of α-helices, which are spirally twisted, are arranged side by side to form long cable like structures. Spiral twisting of the helix introduces an overall left handed twist to the cable itself. This twisting is essential for optimization of the packing of the side chains of the amino acid residues trapped at the interface of the arranged helices. The best optimization of the packing of different α-helical units is achieved when they interact at an angle of about 18^0. The stretchability seen in hair and wool on application of force can be attributed to untwisting of the twist present in the α-keratin. In addition to the hydrogen bonds, the multiple helices in α-keratins may also contain inter-chain disulfide linkages. In fact, the number and pattern of these disulfide bonds determines the extent of curliness in the hair of individuals. Hairdressers use

reduction reactions to break original disulfide bonds and reconstruct them using mechanical and chemical means to generate desired kind of hair curls on demand.

The keratin present in silk contains anti-parallel β-sheets as the secondary-structural motif and accordingly it is referred to as β-keratin. These proteins have an unusual amino acid composition and sequence. The amino acid composition is dominated by glycine, serine and alanine, and every alternate position in the primary structure is occupied by glycine. The side chains of successive amino acids alternatively point upward and downward putting all side chain-hydrogens of glycine residues on one side (let us call it glycine-surface) of the flat β-sheets. Multiple β-sheets are subsequently packed together resulting in to stacked sheets having glycine-surfaces (sometimes even alanine-surfaces) interlocked with each other. Since the β-sheets is an extended structure and these sheets are interlocked, silk is a relatively rigid material that resists stretching. Apart from silk, feathers and scales are the other examples of β-keratin containing tissues.

Collagens

Collagen is the most abundant single protein in most vertebrates and performs a variety of functions. It is an important constituent of bone-matrix, skin and tendons. The amino acid composition of collagens is invariably dominated by glycine, proline and hydroxy proline. The amino acid sequence is generally characterized by recurrence of the tripeptide Gly-X-Pro or Gly-X-hydroxyproline where X can be any amino acid including proline. Presence of large numbers of proline/hydroxyproline residues restrains the polypeptide chain to acquire an α-helical or a β-sheet structure. Instead, individual collagen polypeptides assume a left handed helical conformation with about 3.3 amino acid residues per turn. Three such polypeptides then wrap around one another in right- handed manner with inter-chain hydrogen bonds holding the chains together (Figure 11). This wrapping brings every third residue of each chain to the center of the triple helix that can not accommodate bulky groups but has sufficient space to take in the hydrogen of glycine. The inter-chain hydrogen bonds in the triple helix involving amino protons and carbonyl oxygen of the covalent backbone are the major stabilizing force with some contribution also coming from the hydroxyl groups of the hydroxy-proline residues. Hydroxylation of proline residues is an enzyme catalysed reaction that requires vitamin C (ascorbic acid). Scurvy, the symptoms that indicate vitamin C deficiency, is the result of inadequate hydroxylation of proline residues that leads to weakening of collagen fibers. Lysine amino acid residues that are generally present at the position X are often hydroxylated and offer sites for attachment of polysaccharides.

Multiple collagen triple-helices pack together to form collagen fibers. Each triple helix unit overlaps its neighbor by about 64 nm producing a structural arrangement with remarkable strength. An additional toughness to collagen fibers is imparted by cross-linking of the individual polypeptides via a reaction involving lysine side chains. This cross-linking continues throughout the life that makes collagen more and more brittle and less elastic. This explains the brittleness of the bone and loss of elasticity of skin in older individuals.

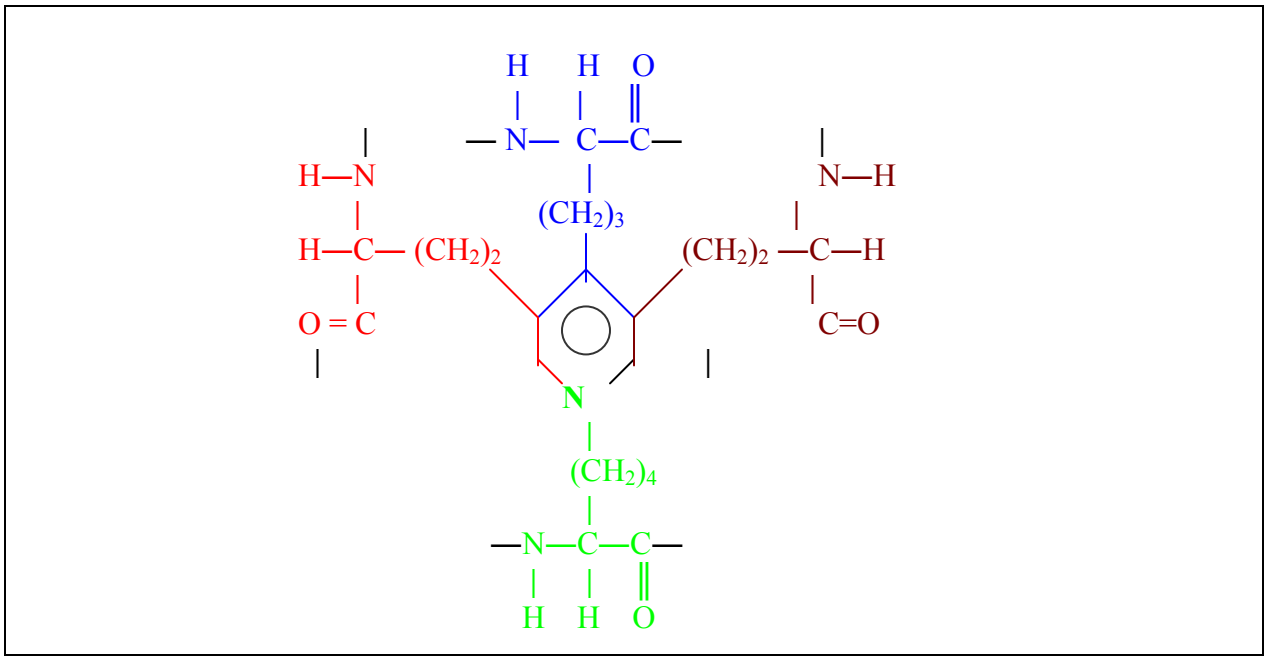
Figure 11: Triple helical structure of collagen.

Elastin

Tissues like ligaments and arterial blood vessels need highly elastic fibers – a property not shared by collagens. Such tissues therefore, contain large quantities of an elastic protein known as elastin. These proteins are very flexible and extendable. They contain large amount of glycine, alanine and valine amino acid residues. They have very little conventional types of secondary structures and generally occur as random coils. The primary structure of these proteins contains significant amount of lysine residues that are probably involved in inter fiber cross-linking. The nature of cross-linking is not similar to what we have seen in the case of collagen. Instead, four lysine residues are interconnected as shown in Box 6. There are fewer cross-links because four polypeptide chains are cross-linked at the same time making the structure highly interconnected rubbery network.

Globular Proteins

Although structural proteins are present in abundance and are essential for architectural purposes, they represent only a small fraction of the kinds of proteins present in the living world. The vital physiological processes including metabolic reactions and transport activities are carried out with the help of an amazing class of proteins known as globular proteins. The polypeptide chains of these proteins are folded into very complex compact and globular conformations. The three dimensional structure of a large number of globular proteins is now known. Two most frequently described globular proteins are myoglobin and haemoglobin. The salient structural features of these two proteins are described in the following section.



Box 6: Interconnection of four lysine residues in elastin.

Myoglobin

Myoglobin is a single polypeptide-chain protein. The total number of amino acid residues and also their sequence varies to some extent from species to species. Human myoglobin, for instance, contains 153 amino acids in its chain. Inspite of minor differences in the amino acid compositions and the primary structures, however, myoglobin serves the same oxygen-storage function in all species. The lone polypeptide chain of myoglobin is folded about a prosthetic group (non-protein, iron-porphyrin complex) called heme. The structure of myoglobin is dominated by helices that are present in eight segments named as A through H. The relative arrangement of these helical segments is depicted in Figure 12. By convention, each amino acid of myoglobin is named after its segment followed by a number indicating its position in that helical segment. For example, the residue number 93 of myoglobin is a histidine that is more often called histidine F8 because it is present at position 8 of the F helical segment of the protein. The iron-binding site of myoglobin resides in its heme group that noncovalently interacts with amino acid residues present in a hydrophobic crevice of the myoglobin. The iron in the heme has six binding positions out of which four are satisfied by nitrogen atoms of the planer porphyrin ring. Of the remaining two coordination sites, which are perpendicular to the porphyrin plane, one is occupied by the nitrogen of histidine F8. Since it is directly interacting with the iron of the heme group, this histidine is called proximal histidine. The sixth coordination site of the heme-iron is occupied either by oxygen or by water (in absence of oxygen). On the second side of the bound oxygen (first side being the heme side) histidine E7 is located which is called distal histidine. The proximity of oxygen to iron is expected to oxidize iron to ferric form. This, however, does not happen as the hydrophobic environment created by the interior of the myoglobin provides a protective antioxidative mechanism. The protective potential of the environment is further enhanced due to sandwiching of the bound oxygen between the heme-iron

and the nitrogen atoms of the distal histidine. This is yet another illustration as to how protein architectural marvels help in running vital physiological processes.

Figure 12: Structure of myoglobin. The picture has been downloaded from the source http://www.ul.ie/~childsp/CinA/Issue64/ToC36-Haemoglobin.htm and adopted after suitable modification.

Haemoglobin

In contrast to the single polypeptide chain present in myoglobin, haemoglobins are known for their subunit structures. Each subunit can be considered to be equivalent to one myoglobin molecule in the sense that all their structural motifs are similar. The normal adult human-haemoglobin has two subunits each of α and β chains and therefore designated as α_2 β_2-tetramer. These chains are themselves very similar to each other structurally and are present in a tetrahedral arrangement. The tetramer can be considered to be composed of two identical dimers, namely $\alpha_1\beta_1$ and $\alpha_2\beta_2$. The two subunits in a dimer are tightly held together but the two dimers can move in relation to each other. The heme groups (see myoglobin), one in each subunits, are very close to the surface but are well separated from each other.

Oxygenation of haemoglobin leads to a significant change in its quaternary structure with minor structural alterations in the tertiary structure as well. The two dimers rotate and slide with respect to each other bringing the two β-chains closer which reduces the size of the central cavity in the molecule as shown in Figure 13. These intra-molecular movements at the level of quaternary structure enable haemoglobin to express an unique characteristic not shared by myoglobin. This property, commonly known as allosteric behaviour, equips haemoglobin to bind oxygen at high concentration in the lung and release it in metabolically active tissues where the oxygen pressure is low.

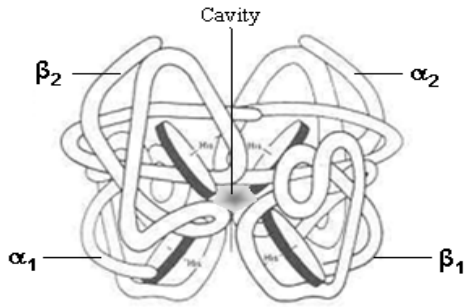


Figure 13: Structure of haemoglobin. The picture has been down loaded from the source http://www.ul.ie/~childsp/CinA/Issue64/ToC36-Haemoglobin.htm and adopted after suitable modification.

Lipoproteins

Owing to their insolubility/poor solubility in aqueous media, lipids (fatty substances) are transported through the blood and lymph in the form of soluble lipid-protein aggregates called lipoproteins. In effect, apoproteins (the protein part of lipoproteins) can be considered as the passive carriers of lipids between the tissues. There are certain apoproteins, however, which perform some additional functions. Apo C-II (Table 3), for instance, act as an activator of lipoprotein lipase, an enzyme involved in triacylglycerol hydrolysis.

Table 3: Classification of human lipoproteins

Type of Lipoprotein (L)	Density (g/ml)	Apoproteins with Molecular Weights	Protein Contents (% Dry Weight)	Diameter (nm)
Chylomicron	< 0.95	A-I (28,300), A-II, B-48 (241,000), C-I, C-II (10,000), C-III	2.00	500-1000
VLDL (Very Low Density)	0.950-1.006	B-100, C-I (7,000), C-II, C-III (9,300), E	8.00	30-70
IDL (Intermediate Density)	1.006-1.019	B-100, C-I, C-II C-III, E (33,000)	15.00	25-30
LDL (Low Density)	1.006-1.063	B-100 (513,000)	22.00	20-25
HDL (High Density)	1.063-1.210	A-I, A-II (17,400), C-I, C-II, C-III, D (35,000), E	40-55	10-15

Depending on the types of lipids and the apoprotein composition of the aggregates, lipoproteins vary in their densities and are classified accordingly (see Table 3). Despite their differences in lipid and protein compositions, there are striking similarities in their structural features. They have spherical shapes with hydrophobic moieties of the lipid and protein components forming its inner core while the hydrophilic components present at the surface (Figure 14). The primary structure of most of the apoproteins is known. The amino acid sequence has certain distinct regions that are rich in hydrophobic amino acid residues that facilitate the binding of apoprotein components to the lipid part of lipoproteins.

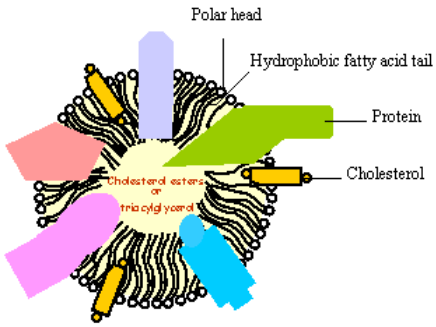


Figure 14: Structure of lipoprotein.

Metalloproteins

Many proteins contain metal ions as their non-amino acid structural constituent and are known as metalloproteins. These metal ions are usually linked directly to the side chains of some amino acids by coordinate bonds. For instance, Zn^{++} ion is linked to His 69, His 196 and Glu 72 amino acid residues of the enzyme carboxypeptidase A. Some times metal ions are structural part of certain prosthetic groups that are attached to the proteins (iron containing heme of myoglobin has already been discussed). Structural features of metalloproteins are generally the same as any other protein. Metal ions attached to the polypeptides play vivid kind of roles and no generalization can be made to this effect. Zinc ion in carboxypeptidase A, for example, is directly involved in the catalytic process whereas iron provides a site for the binding of oxygen in myoglobin. Copper ion present in ascorbic acid oxidase enzyme is involved in oxidation-reduction reaction. Similarly, cobalt and nickel ions are structural constituents of the cobalamine (coenzyme for the enzyme glutamate mutase) and the urease enzyme respectively.

Glycoproteins

Over fifty percent of proteins present in eukaryotic systems contain covalently attached carbohydrate chains of varying lengths. These proteins are referred to as glycoproteins. There is another related term called proteoglycans which simply refers to a situation where the relative content of the carbohydrate far exceeds (over 90 % by weight) the protein content of the protein-carbohydrate conjugate. Those proteoglycans that contain small peptides in place of proteins are called peptidoglycans.

The protein part predominates in glycoproteins as their carbohydrate components are generally short and branched chains of about 15 residues or less. If the saccharide chain is linked to proteins through a linkage between N-acetylglucosamine/ N-acetylgalactosamine of the sugar and the side-chain amino group of asparagines of the protein component, it is called N-linked glycans. They are named O-linked glycans when the glycan part is linked to the hydroxyl group of a threonine or serine residue of the protein (Figure 15). Occasionally, the side chains of hydroxyproline residues are also employed for these kinds of linkages (for example collagen). Blood group antigens offer the best examples of O-linked glycans. Human immunoglobulin G and hen ovalbumin contain N-linked glycans.

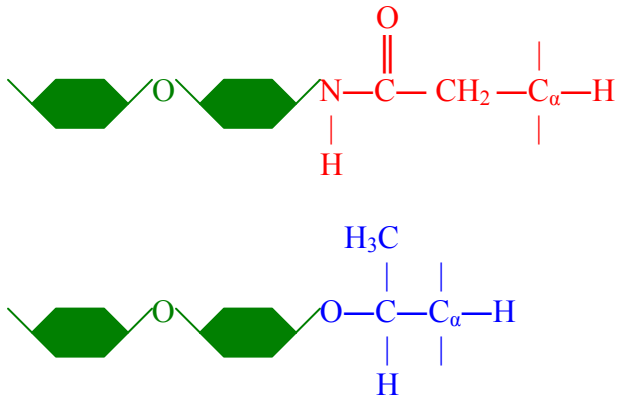


Figure 15: Two ways of linking oligosaccharide chains (shown in green) through the amide nitrogen (N-linked glycans) of asparagine residue (shown in red) and the hydroxyl oxygen (O-linked glycans) of threonine (or serine) residue (shown in blue) of proteins

What kinds of role these carbohydrates exactly play in the functioning of glycoproteins is still a subject of speculation. However, it is believed that oligosaccharide chains guide glycoproteins to their intracellular destinations. They provide stability to the structure and are likely to help in acquisition of three-dimensional structure of proteins. In addition to their role as an antigen on the cellular surfaces that has led to blood grouping, they have also been viewed as structural components in biological systems.

Nucleoproteins

Nucleic acids interact with proteins to form complexes collectively known as nucleoproteins. Association of proteins to nucleic acids may be transient in nature as happens during course of nucleic acid metabolism or it may lead to the formation of stable complex as seen in chromatin. Nucleoproteins are major structural constituents of ribosomes and accordingly these proteins are called ribonucleoproteins. Ribosome is held together by ribosomal proteins that interact with themselves on one hand and with the ribosomal RNA on the other. Interaction of ribosomal RNA with proteins helps in the folding of the latter while the same interaction protects the former from being degraded by nucleases.

Nucleoproteins are also present in the chromatin of the nuclei. These proteins have been classified as histone proteins and non-histone proteins. Histones are of five types with molecular weight varying between 11.2 to 22.5 kDa. All histones are highly basic in character and are rich in arginine and lysine. Owing to their basic character, histones are ideal choice for stabilizing highly acidic DNA having large number of negative charges. The histones are thus considered the basic building blocks of chromatin structure.

In contrast to histones, non-histone nucleoproteins have a diverse nature. Their number at about 1000 is far more than the histones but they are present at relatively much lower concentration. These proteins include different nuclear enzymes, regulatory proteins and various hormone receptors.

Proteins in Health and Disease

Human body requires all the twenty amino acids for the synthesis of its proteins. Half of these amino acids (Arg, His, Ile, Leu, Lys, Met, Phe, Thr, Trp, Val) can not be synthesized in the body in adequate amounts and must be provided in the diet. These amino acids are therefore called essential amino acids. The remaining amino acids are called non-essential because the body can synthesize them from the available precursors. Typical mixed diets contain proteins that provide ample amounts of both essential and non-essential amino acids. Based on their content of amino acids, food-proteins are classified as complete, partially complete or incomplete proteins. A complete protein contains enough of all the essential amino acids. Egg (ovomucoid and ovalbumin), milk (casein) and meat (myosin) proteins are examples of complete protein and are said to have very high biological value. Partially complete proteins maintain life, but they contain inadequate amounts of some of the essential amino acids. The wheat protein, gliadin, is a notable example of this class of protein. Totally incomplete proteins are incapable of replacing or building new tissues. Zein, the corn protein, and gelatin are classic examples of totally incomplete proteins.

A protein intake that fails to meet the body requirement leads first to depletion of tissue reserves and then to a lowering of the blood protein levels. Nutritional edema is a clinical sign that appears after substantial depletion of tissue reserves. *Kwashiorkor* (meaning "the displaced child), a disease characterized by failure of growth, skin leisions, edema and change in hair colour, is caused due to lack of good quality protein in the food. Kwashiorkor is generally accompanied with a disease called Marasmus which is caused due to an insufficient intake of calories. The dominant characteristics of Marasmus include loss of subcutaneous tissue and

muscle atrophy. Mixed protein-calorie deficiency presents symptoms of Kwashiorkor – Marasmus.

Chemical Synthesis of Polypeptides

We have seen on page 5 that a peptide bond could be formed between two amino acids by elimination of a water molecule. This reaction is, however, not so simple as the free energy change for it is + 10 kJ/mol which means that the reaction is not energetically favoured. In the cell, peptide bonds are formed in a complex manner using the cellular machinery. For synthesizing peptides and proteins chemically, the reaction is first made energetically favourable. Since amino acids have multiple reacting groups, the amino and carboxyl groups that are to remain unlinked and also the side chains will have to be protected from undergoing undesired reactions (Figure 16). This is achieved by protecting the NH_2 group of one amino acid (say AA_1) by blocking it with a suitable group like carbobenzoxy, tosyl, trifluoro or t-Butyloxycarbonyl (t-Boc). Likewise, the carboxyl group of another amino acid (say AA_2) is protected by chemically blocking it with ethyl, t-Butyl or nitrobenzyl group. The carboxyl group of AA_1 is then activated with suitable reagents like dicyclohexylcarbodiimide (DCC) or mixed anhydrides. A reaction between the activated carboxyl group of AA_1 and amino group of AA_2 leads to the formation of a peptide bond between the two amino acids. The resultant dipeptide, AA_1-AA_2, still have blocked NH_2 and COOH termini that are freed by exposing it to mild acidic conditions (Figure 17). All excess reagents and the byproducts are removed yielding pure dipeptide.

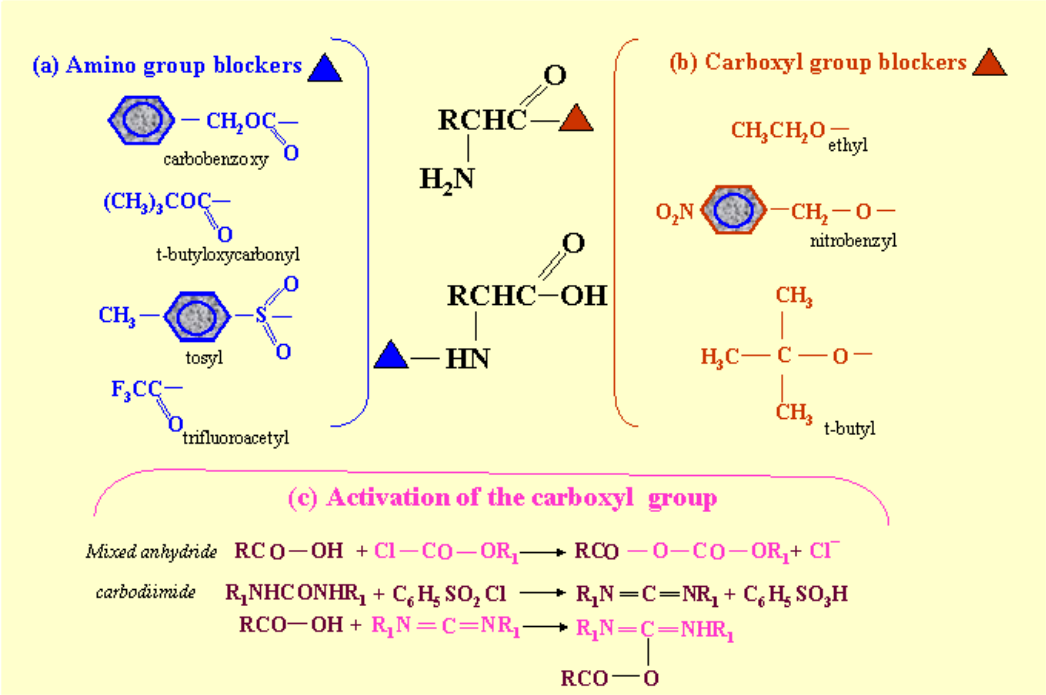


Figure 16: Chemical groups used for protection of amino (a) and carboxyl (b) groups; and for activating the carboxyl group (c) in chemical synthesis of peptides.

For synthesizing a tripeptide, the dipeptide (with blocked amino group) obtained above is allowed to react with the third amino acid (with blocked carboxyl group) after activating its carboxyl group as described above. The process can be repeated several times to synthesize the required peptide. This method, however, has serious limitations because complete removal of potentially contaminating reaction products during successive cycles is a very cumbersome process that becomes increasingly difficult with increase in number of cycles.

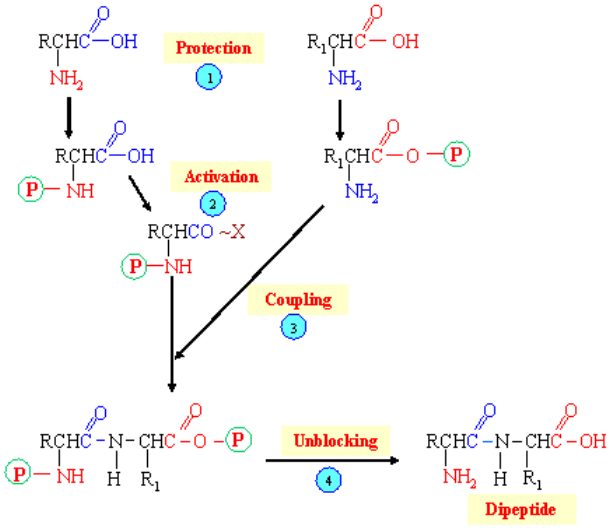


Figure 17: Schematic representation of peptide synthesis. Non-reacting groups of the amino acids to be linked are first protected (Step 1). The unprotected carboxyl group is subsequently activated (Step 2) and then allowed to react with the unprotected amino group of the other amino acid (Step 3). Protecting groups are finally removed from the product (Step 4).

Merrifield Solid-phase Peptide Synthesis

The method for peptide synthesis described above is very cumbersome as removal of excess reagents and byproducts is required at every step. Based essentially on the same principle as described above, R. Bruce Merrifield developed a neat and simple procedure for the chemical synthesis of peptides for which he was awarded Nobel Prize in 1984. The solid-phase method – as it is generally called, has a big advantage because the desired product at each step is bound to solid polystyrene beads that can be easily filtered and washed resulting into good yield with high purity. All amino acids that are to be incorporated in the desired peptide are treated with t-Boc to protect their amino groups. The t-Boc derivative of the carboxyl terminal amino acid residue of the perspective peptide is then anchored to polystyrene beads (through the carboxyl group of the amino acid) kept in a vessel. The protecting t-Boc is subsequently removed which liberates amino group of the amino acid free. The t-Boc derivative of the next amino acid is then added to the vessel together with the coupling agent, dicyclohexylcarbodiimide. After formation of the peptide bond, excess reagent and the byproduct, dicyclohexylurea, are washed away leaving behind the dipeptide anchored with the beads. Remaining amino acids are linked into the peptide chain by repeating the same sequence of reactions. At the end of the peptide synthesis, the peptide is detached from the beads by treating it with hydrofluoric acid that selectively cleaves

the carboxyl ester bond that anchors the peptide to the beads (Figure 18). The whole process of the solid-phase synthesis has been automated and peptides containing up to about 100 amino acid residues can be synthesized in good yield with a high level of purity.

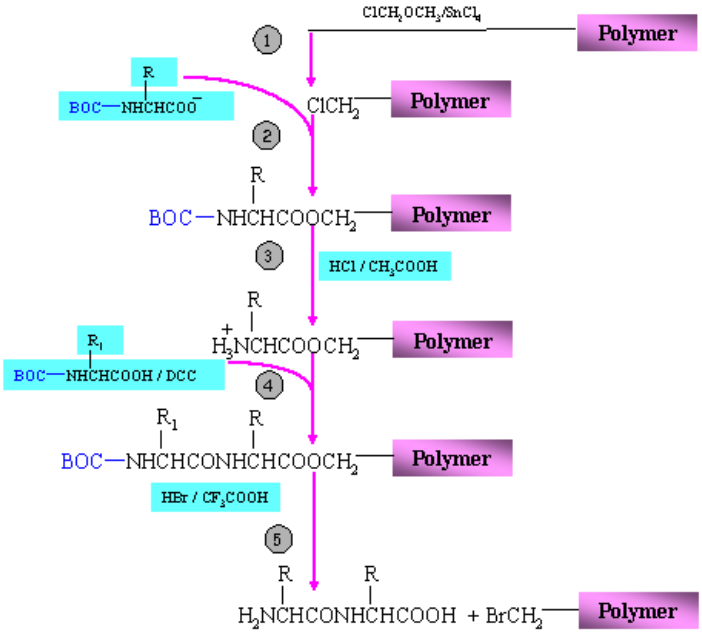
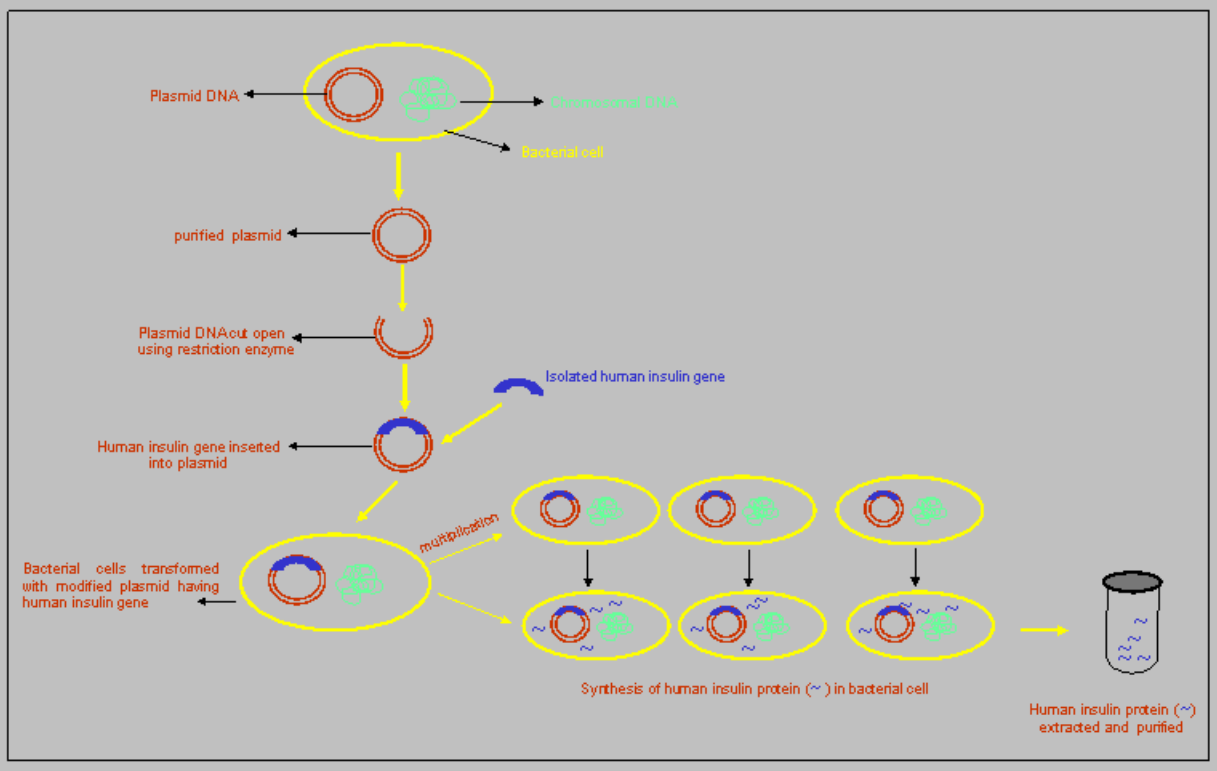


Figure 18: Merrifield solid phase synthesis of peptides. Different steps include (1) activation of the solid polymer support (2) linking of the t-butoxycarbonyl (BOC) blocked amino acid to the activated polymer (3) removal of the BOC group (4) linking of the second amino acid with the BOC blocked amino and dicyclohexylcarbodiimide (DCC)-activated carboxyl group to the polymer - bound first amino acid and (5) the release of the peptide from the polymer along with the removal of the BOC protecting group by hydrogen bromide and triflouroacetic acid treatment.

Protein synthesis by recombinant DNA technology

Recently, the technique of recombinant DNA technology has been successfully used to synthesize proteins using cellular conditions. Many clinically and commercially important proteins are being produced by inserting the gene of interest for the respective protein into a suitable plasmid/vector. These recombinant plasmids are then introduced in bacteria to produce the protein that is chemically identical to its naturally produced counterpart. The use of recombinant DNA technology in the commercial production of human insulin is given in the form of a flow chart in Box 7. In the very first step, plasmid DNA from bacterial cells is isolated, purified and cut-open with suitable restriction enzymes. The human insulin gene is inserted into the open plasmid and ligated using ligase enzyme. The plasmid is then introduced in bacterial cells which are thus transformed with this recombinant plasmid having insulin gene. Transformed bacteria are allowed to multiply producing millions of copies the insulin gene that eventually translates into the insulin protein. The insulin is finally extracted and purified from the bacterial cell for further use.

BOX 7: A general approach of recombinant DNA technology showing production of human insulin protein in bacterial cells.



Determination of Amino Acid Sequence of a Polypeptide Chain

Amino acid sequence (primary structure) of a protein can be determined either by using a multi-step chemical procedure or by finding out the sequence of the gene that codes for it. The chemical procedure is a long process and needs a pure sample of the protein to be sequenced. The multiple polypeptide chains of the protein, if present, are separated and each chain is sequenced separately. The major steps of sequencing of a polypeptide chain by chemical method are listed below.

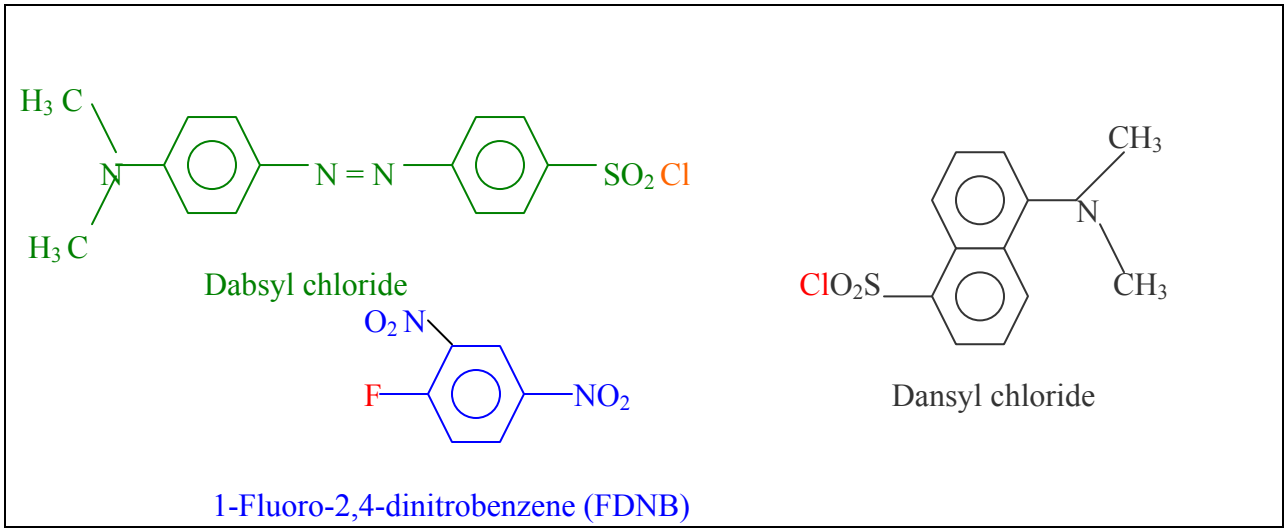
1. Determination of the molecular weight of the polypeptide chain
2. Determination of amino acid composition the polypeptide chain
3. Identification of the amino and carboxyl termini of the polypeptide
4. Preparation of multiple sets of fragments (small peptides) of the polypeptide chain by specific enzymatic and chemical cleavages
5. Determination of molecular weight and amino acid composition and identification of the amino and carboxyl termini of all fragment obtained at step 4.
6. Sequencing of the fragments obtained at step 4.
7. Construction of the protein sequence from the data obtained at step 6 by obtaining overlapping sequences in the primary structures of fragments.

Information on the molecular weight is essential for expressing the results in molar terms if the sequencing is started with known amount (by weight) of the protein. Molecular weight can be determined by a number of physical techniques including osmotic pressure, ultracentrifugation, gelfiltration and sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDSPAGE).

SDSPAGE, a preferred technique by many, is based on mobility of proteins in an electrical field in presence of the detergent, sodium dodecyl sulphate. The mobility of proteins becomes independent of their shape and charge and depends only on their molecular weight because they acquire linear shape in the presence of the detergent. By comparing the relative mobilities with standard proteins of known molecular weight, the molecular weight of the unknown proteins/peptides can be accurately computed.

The number of each amino acid residue present in one molecule of the protein (amino acid composition) can be determined after hydrolyzing all peptide bonds in 6 N HCl under vacuum or in presence of N_2 in a sealed tube. The tube is kept at about 110 ^{0}C for over 24 hours. The protein hydrolysate thus obtained contains most of the constituent amino acids in free form. The acid is evaporated off and the residue solubilized in an appropriate buffer. The amino acids are finally separated and quantitated by a machine called amino acid analyzer which works on the principle of ion exchange chromatography. The quantitative recovery of a few amino acids from the acid hydsrolysate is not possible. For instance, tryptophan is completely destroyed during the acid hydrolysis and it is generally quantified from alkali hydrolysates of proteins. Methionine is also sensitive to acid hydrolysis and it can be quantitatively recovered by inclusion of adequate amounts of β-mercaptoethanol in the hydrolysis mixture. Similar strategic treatments may also be required for the quantification of cysteine, asparagene and glutamine residues of proteins. The information on the amino acid composition is used for the verification of the data obtained on the sequence of the protein.

Knowledge on the identity of the two termini of the protein is also required to deduce its sequence. The carboxyl terminal end of a protein can be identified by treating it with specific enzymes called carboxypeptidases. These enzymes successively release amino acids from the carboxyl terminal end of the protein. By closely monitoring the rate and the order of release of amino acids, the carboxyl terminal amino acid residue of the protein can be identified.



Box 8: Some common reagents used for identification of amino terminal residues of proteins

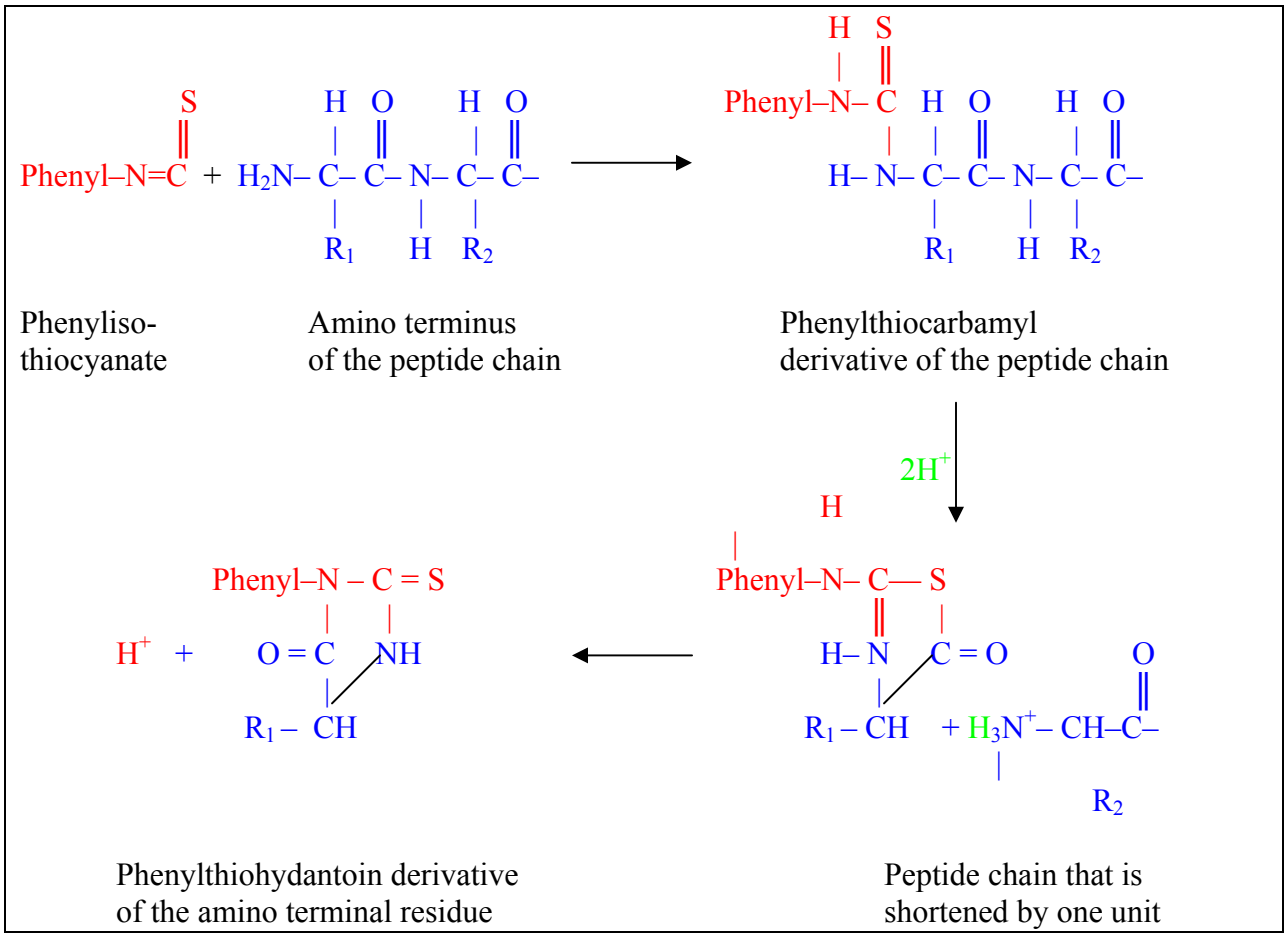
The amino terminal amino acid residue of a protein/peptide can be identified by labeling it with compounds which form stable covalent links. Frederick Sanger was the first person to use 1-fluoro 2,4-dinitrobenzene (FDNB) for this purpose. Owing to its poor sensitivity, FDNB is not in much use these days. Another compound called dabsyl chloride or a related compound known as dansyl chloride is now commonly used for labeling amino groups as they offer a sensitive alternative to FDNB. All these labeling compounds (Box 8) reactive moieties are shown in red) form stable derivatives with the α NH_2 group of the protein. Acid hydrolysis of the protein yields the amino terminal residue as labeled amino acid that can be identified by chromatographic techniques.

Although some of the above described methods are very simple and sensitive, they can not be used repeatedly on the same protein/peptide because the latter is totally degraded during the acid hydrolysis. A compound called phenyl isothiocyanate was therefore introduced by Pehr Edman for labeling amino terminal with unprecedented advantage. Phenyl isothiocyanate reacts with the amino group to form phenylthiohydantoin (PTH) derivative that can be hydrolytically removed under mild acidic conditions without affecting the rest of the protein/peptide chain that is now shortened by one amino acid. The liberated derivative can be identified chromatographically. The Edman's procedure can again be repeated on the shortened protein to identify the next amino acid of the sequence (Box 9). Although, this process can be repeated a number of times but it becomes more and more difficult as the number of cycles increases. Thus, in addition to identification of the amino terminal amino acid residue, the Edman procedure can be used for sequencing of small peptides but it can not sequence large protein polypeptide chains. The applicability of the method has, however, increased considerably after the development of machines called sequenators that work on this very principle.

Fragmentation of Proteins

By Specific Enzymatic and Chemical Reactions

Since only shorter peptides comprising fewer than about 50 residues can be reliably sequenced by Edman method, the long protein polypeptides are cleaved in to smaller peptides to facilitate their sequencing. Specific cleavage can be achieved by chemical or enzymatic means. Bernhard Witkop and Erhard Gross, for instance, have found out that cyanogens bromide (CNBr) splits polypeptide chains specifically on the carboxyl side of methionine residues i.e. only those peptide bonds are cleaved whose carbonyl function is contributed by methionine. Thus, a protein containing two methionine residues would yield three fragments on its fragmentation with CNBr. The size of these fragments would depend on the position of the methionine residues in the primary structure of the protein. Proteins can also be subjected to controlled enzymatic hydrolysis by proteolytic (protein degrading) enzymes. For example, trypsin specifically cleaves only those peptide bonds whose carbonyl function is contributed by either a lysine or an arginine amino acid. A protein having two lysine and one arginine residues would thus generally give four fragments on its treatment with trypsin. Several other cleaving agents with their specificity requirements are listed in Table 4.



Box 9: The Edman degradation procedure of sequencing small proteins/peptides

Table 4: Specificities of some protein cleaving agents

Agents	Nature	Specificity
Trypsin	Enzyme	Carboxyl side of arginine and lysine residues
Chymotrypsin	Enzyme	Carboxyl side of tyrosine, tryptophan and phenyl alanine
Clostripain	Enzyme	Carboxyl side of arginine residues
O-Iodosobenzoate	Chemial	Carboxyl side of tryptophan residues
Cyanogen bromide	Chemical	Carboxyl side of methionine residues
2-Nitro-5-thio-cyanobenzoate	Chemical	Amino side of cysteine residues

Fragments present in the protein hydrolysates obtained after its treatments with aforesaid agents can be purified by standard separation techniques. The techniques generally employed for this purpose include gel filtration, ion exchange chromatography, high-performance liquid chromatography, affinity chromatography and ultracentrifugation.

Gel Filtration

Separation of peptides/proteins by gel filtration is based of their molecular size. A column is packed with porous gel beads made up of highly hydrated polymers like dextran, agarose or polyacrylamide. The porosity of the gel is strategically chosen so that the smaller molecules of the mixture can enter the beads leaving the larger ones in the interstitial space. Once the sample is loaded on top of the column, the smaller molecules move slower for they spend a part of their time in flowing through the gel beads whereas the lager molecules only flow through the interstitial space. If fractions are collected as the solvent is passed through the column, the earlier fractions will contain the larger molecules leaving the smaller ones behind (Figure 19).

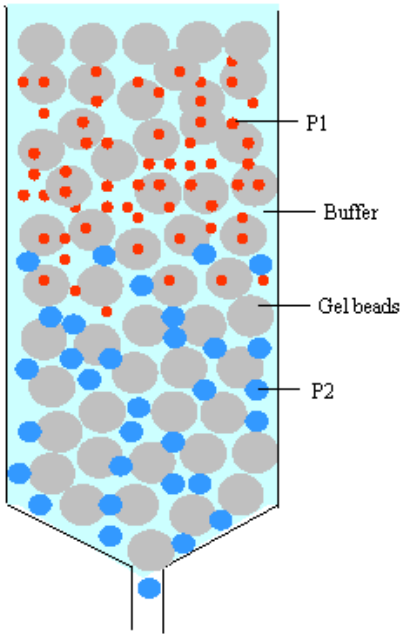


Fig 19: **Separation of proteins by gel filtration**. The constituents of the mixture containing proteins P1 (small in size, shown in red) and P2 (larger in size, shown in blue) getting separated on a gel filtration column on the basis of their sizes.

Ion-exchange chromatography

Separation of proteins/peptides by ion-exchange chromatography is based on the net charges on them. Chromatographic column is filled with the chromatography media which is generally a synthetic resin or cellulose containing charged groups. The media possessing positive and negative charges are called anion- and cation-exchangers that have the potential of binding proteins having net negative and positive charges respectively. Charged proteins bound to such chromatographic columns can be eluted /unbound by increasing the concentration of salt in the eluting buffer. Proteins that have a low charge density (weaker binding to the column) will tend to emerge first, followed by those having a higher charge density (Figure 20).

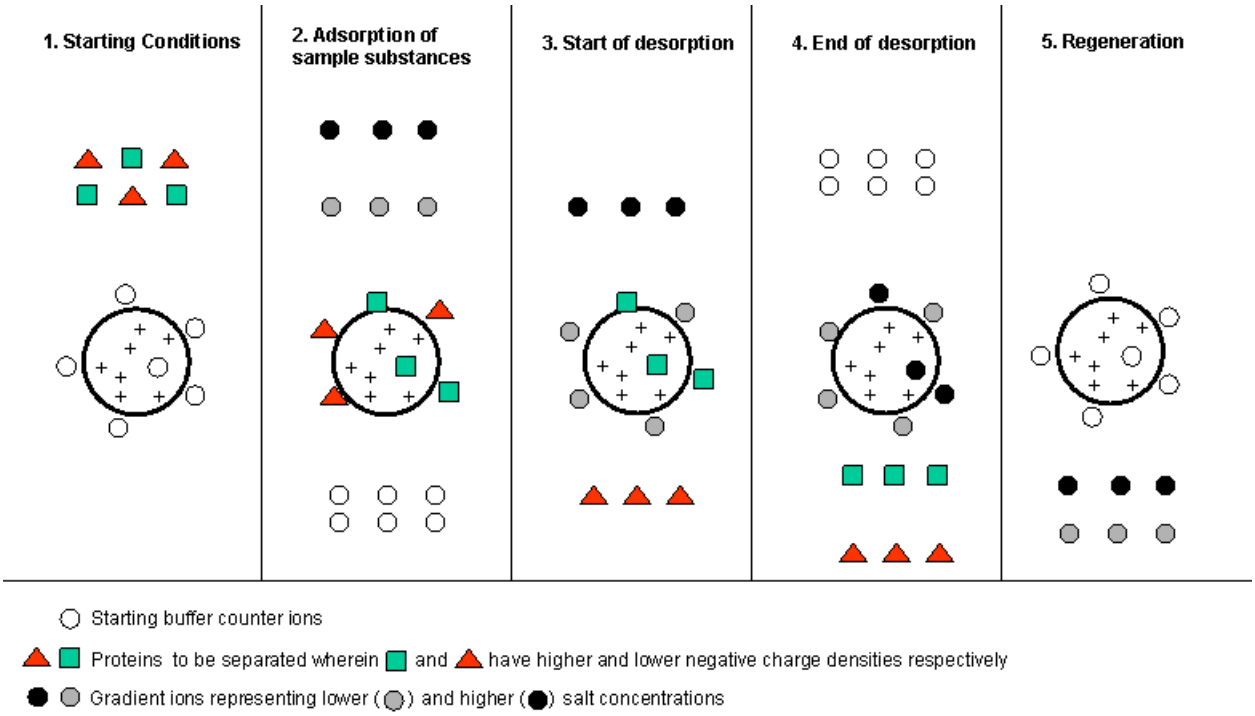


Fig 20: Separation of negatively charged proteins by ion exchange chromatography on an anion exchanger.

High performance liquid chromatography

High-performance liquid chromatography is not so much a new type of chromatographic technique as a modern way of enhancement of the resolving potential of the old techniques. Principles are the same but the chromatographic media consist of finely divided particles with improved apparatus designed to permit chromatography at high pressures allowing better separations at faster pace.

Affinity chromatography

Separation by affinity chromatography is based on the high affinity of many proteins for specific molecules/groups generally called ligand. The ligand is covalently coupled to a suitable

chromatographic media in such a manner that it still retains its characteristic protein binding property. As the protein mixture is allowed to pass through the column, the protein having affinity for the immobilized ligand binds to the column, while other proteins pass through.

Ultracentrifugation

Molecular weight, shape and size of protein molecules govern their separation by ultracentrifugation. A small volume of the protein solution in a cell is carefully subjected to controlled centrifugal force by the ultracentrifuge machine that operates above 50,000 rpm. The initially uniformly distributed solute molecules start moving towards the bottom of the cell. This movement of protein molecules forms a boundary between the solute depleted solvent region at the top and the solution in the lower region of the cell**.** Since different constituents of the mixture sediment at different velocity, boundaries involving different constituent appear with progression of the centrifugation process (Figure 21). Movement of boundaries can be monitored with an optical device and they can be separated by suitable mechanical devices.

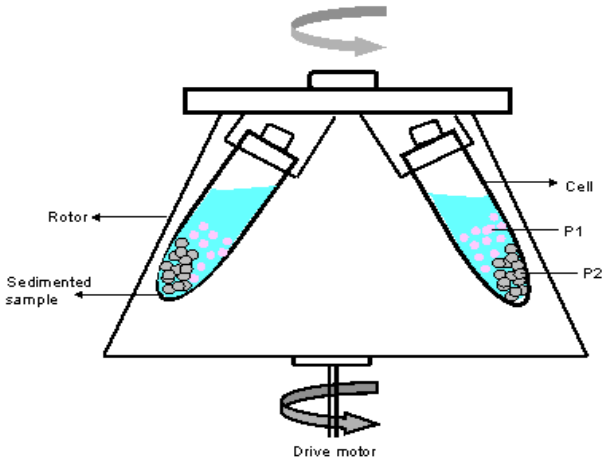
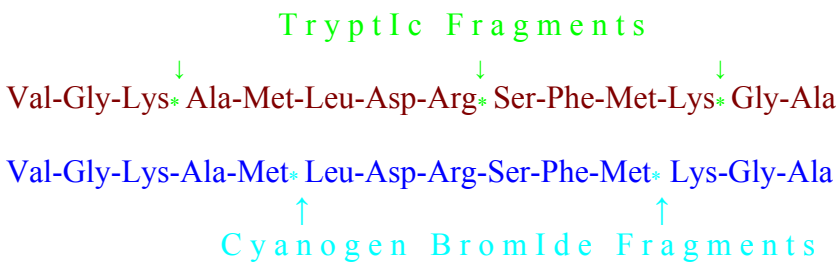


Fig.21: **Separation of proteins by ultracentrifugation**. The constituents of a mixture containing proteins P1 (low molecular weight, shown in pink) and P2 (higher molecular weight, shown in grey) sedimenting at different rates.

Sequencing of the Fragments/Peptides and Construction of the Amino Acid Sequence of the Complete Protein Chain

Multiple sets of small fragments/peptides of the target protein are prepared by splitting it with more than one cleaving agents. The purified fragments are characterized in terms of their molecular weight, amino acid composition and the amino- and carboxyl-terminal residues as has been described for the intact protein above. The amino acid sequence of each of these fragments is determined by the Edman method. The sequence data on each set of fragment would independently represent the amino acid sequence of different segments of the protein. However, the order of these segments in the complete protein chain is still unknown. This order of the segments (fragments) can be determined by obtaining peptide overlaps of the multiple sets of fragments as illustrated in Box 10.

If the amino acid sequences of the first-set of three fragments obtained by CNBr cleavage of a protein are: Leu-Asp-Arg-Ser-Phe-Met, Lys-Gly-Ala and Val-Gly-Lys-Ala-Met, and the sequences of the four fragments of the *second-set* obtained by tryptic cleavage are: Ser-Phe-Leu-Lys, Gly-Ala, Val-Gly-Lys and Ala-Met-Leu-Asp-Arg, the two sets of fragments can be arranged to overlap each other as follows:



Keeping in view the identity of the amino and carboxyl terminal amino acid residues of the intact protein and the overlapping pattern of the two sets of the fragment depicted above, the complete amino acid sequence of the polypeptide chain can be deduced as: Val-Gly-Lys-Ala-Met-Leu-Asp-Arg-Ser-Phe-Met-Lys-Gly-Ala.

Box 10: Deduction of the amino acid sequence of a protein

The method for protein sequencing described above can be applied to any protein having a single polypeptide chain and devoid of intra-chain disulfide bonds. Proteins having disulfide bonds and also those containing multiple subunits require additional experimentation for their sequencing. All polypeptides of multisubunit-proteins lacking disulfide linkages are separated by treating them with suitable denaturing agents like urea or guanidine hydrochloride. Subunits linked with disulfide bonds are separated after reducing these linkages with chemicals like dithiothretol or β-mercaptoethanol followed by the treatment with a denaturing agent. Reoxidation of cysteine residues is stopped by their alkylation with an alkylating agent like iodoacetic acid.

Once separated, these polypeptides can be sequenced in usual manner. The positions of disulfide bonds can be ascertained by a technique called diagonal electrophoresis. The protein is specifically cleaved into a set of fragments under conditions in which the disulfide bonds are not

affected and remain intact. Clearly, all peptides would be released in a free state except those pairs of peptides that contain the disulfide linkages. This mixture of peptides is then subjected to paper electrophoresis along one dimension where all peptides would separate on the basis of their charge to size ratio. The paper sheet is then exposed to performic acid vapours, which oxidizes disulfide bonds to cysteic acids. The electrophoresis is again performed in second dimension under the same experimental conditions. Peptides that were devoid of disulfides will move exactly the same way as they did earlier and therefore they would align on a single diagonal line as shown in Figure 22. However, the electrophoretic mobility of the newly formed peptides (shown by red squares in Figure 22) with cysteic acid would be different from their parent peptide and hence will not lie on the diagonal line. The location of these cysteic acid residues in the sequence would thus indicate the location of the disulfide bonds.

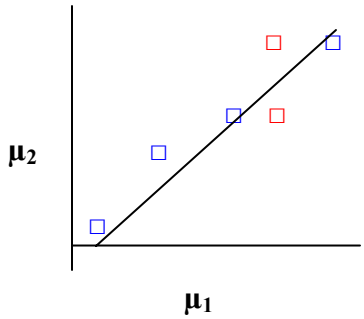


Figure 22: Assignment of positions of the disulfide bonds in their primary structure. Paper electrophoresis of the hydrolytic mixture of the protein is performed where constituent peptides will move with their respective electrophoretic mobility, μ_1. The electrophoretogram (i. e. the paper retaining the protein fragments after first round of electrophoresis) is then sprayed with performic acid that oxidizes disulfide bonds to cysteic acids. The electrophoresis is again performed in the second dimension under similar conditions. Peptides move again with their electrophoretic mobility, μ_2, which is the same as μ_1 for all peptides except the ones freed by the performic acid treatment. All peptides (shown by blue squares) would thus align themselves diagonally as shown in the figure except the ones that were linked by disulphide bonds (shown by red squares)

An alternate strategy for protein sequencing, which has become popular in recent times, involves the sequencing of the gene that codes for the protein under reference. Since the information on the amino acid sequence of a protein resides in its m-RNA (in the form of triplet codes) which in turn takes this information from the gene (a DNA segment), the sequencing of the latter directly gives the sequence of the protein. Availability of relatively faster and accurate methodology for DNA sequencing has made this procedure of protein sequencing very popular among the scientists.

Suggested Readings

1. Biochemistry, 3rd edition, by Christopher K. Mathews, K. E. van Holde and Kevin G. Ahren.
2. Biochemistry, 5th edition, Lubert Stryer, Jeremy M. Borg, and John L. Tymoczko.
3. Lehninger Principles of Biochemistry, 4th edition, by David L. Nelson and Michael M. Cox..
4. Textbook of Medical Biochemistry, 2nd edition, M.N. Chatterjea and Rana Shinde Published by Jaypee Brothers Medical Publishers (P) Ltd., New Delhi