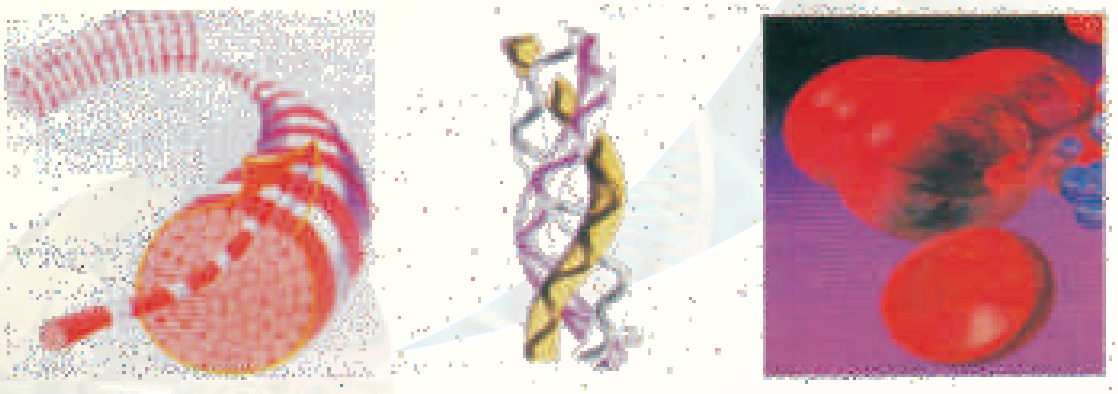
CHAPTER

2

PROTEIN STRUCTURE AND ENGINEERING

5.2.1. Introduction to the World of Proteins

The shape, structure and function of the human body is one of the Nature's marvels. The fertilisation of an egg by a sperm to the growth of a whole human body involves numerous steps of growth and differentiation. When we breathe, we feel a sense of oxygen flowing through our lungs and racing in our blood vessels, to be delivered to all our tissues. While we flex our muscles, we can feel them first tightening and then relaxing. The molecules involved are proteins, haemoglobin which transports oxygen, collagen which provides the strength to our bones and extracellular tissue and actin, myosin and several others which help in muscle contraction **(Fig. 1)**. It is noteworthy that among the biomolecules you have studied, proteins have the maximum diversity in function. The key to this enormous diversity is the unique structure of proteins. Although all proteins are made up of 20 different amino acids the sizes and sequence combinations and variations of each protein leads to millions of unique 3-D structures and thereby functions. Scientists have been striving to relate protein structure with function and hence the first step would be to determine 3-D structure of a protein.



Myosin and actin | Collagen | RBC with Haemoglobin

Fig. 1. Proteins having multiple roles

Even more amazing is the structure and data processing abilities of the human brain. We tend to marvel at the incredible speed and processing functions of the super computers little realising the creativity of the human being who invented them. The speed and correlation of sensory stimulation is unique to the human brain which can grasp the diversity of sensory inputs and convert them to learning and memory for later application. What are these proteins which enable

these functions and why are some brain related diseases like Alzeimers occurring, in which certain proteins show abnormal structure and behaviour?

A number of human diseases are due to the deficiency or abnormal structure of proteins. The lack of a particular subunit, alpha or beta of the oxygen carrying protein haemoglobin results in Thalassaemia, a devastating disease in which an infant cannot grow without repeated transfusions. If the beta chain is present but with a substituent in one of the amino acid residues another debilitating condition called Sickle Cell Anaemia results which is endemic to certain parts of Africa. The absence of an enzyme- Adenosine deaminase results in the birth of a severely immunocompromised baby who cannot last infancy (SCID). More recently it has been discovered that certain "rogue proteins" whose structure has been altered can result in diseases such as the Mad cow disease wherein the disease itself appears to be propagated by infectious proteins called **"prions"**. Clearly proteins need to be understood in detailed terms.

The completion of the Human Genome Sequence has revealed about 35,000 genes. However the actual number of proteins encoded by these genes may be many more due to posttranscriptional modifications. Different cells have specialised proteins for their unique functions in addition to the housekeeping proteins required for metabolism and generation of ATP. Sometimes these proteins are secreted to the outside like the proteolytic enzymes from the pancreas or hormones from ductless glands like the pituitary. We are yet to identify all the proteins required for a body to function and this presents a challenge to the future biotechnologists. One of the outcomes is the merging field of protein structure and function- proteomics. This chapter will enable you to understand various features of the area of proteomics- 3-D structure, functions and applications of protein products, some generated by biotechnological processes.

5.2.2. 3-D Shape of Proteins

The morphology, function and activity of a cell are all dependant on the proteins expressed. Proteins perform a variety of roles. The three dimensional properties of proteins have an important bearing on their function. The first step in determining the structure of a protein is to isolate it in a pure form from its cellular location (note bacterial, plant or animal cell). The purified protein is then crystallised so that using a technique called X-ray crystallography its three dimensional structure can be deduced. Nowadays another powerful technique called Nuclear Magnetic Resonance (NMR) has been developed which can deduce protein structures in solution and hence crystallisation is not required. However the protein in either technique has to be purified and some general procedures used to purify proteins will be discussed in subsequent sections of this chapter. In general when we refer to the structure of a protein this involves two aspects- the chemical structure which is the amino acid sequence of the polypeptide and its folding in space which is referred to its 3-D structure.

One of the major breakthroughs in protein sequence determination was achieved in the middle of the last century by Dr. Frederick Sanger who developed the first sequencing reagent FDNB (fluoro dinitro benzene) and a general strategy for sequencing. By using these methods he was able to sequence the important hormone insulin which is required by diabetics and more importantly he demonstrated for the first time that proteins were linear polymers of amino acids. For this work he was awarded the Nobel Prize and it will be interesting for you to know that several years later he was awarded a second Nobel Prize for developing a sequencing technique for DNA which has been described in the Recombinant DNA technology chapter previously. Another protein chemist, Pehr Edman in 1950 developed another sequencing reagent and procedure which is used in modern day sequenators as the procedure has been automated. These methods have been discussed in the XIth class' textbook of Biotechnology in Unit II, Chapter 2. Notably using the sequence of insulin established by Sanger a biotechnology company called Eli Lilli was able to develop recombinant human insulin which is the major source for insulin administration to diabetics worldwide.

With the availability of pure proteins, scientists like Linus Pauling, G.N.Ramachandran, Max Perutz and John Kendrew to name a few started developing techniques to study the 3-D shapes of proteins using high resolution X-rays. They laid the foundation for deducing protein structure by enunciating the basic rules which govern protein folding and the forces which cause the folding and stabilise them. Hence from these studies the concepts of planarity of the peptide bond, secondary structures such as alpha helix and beta pleats were developed. These concepts were introduced in the Class XI Biotechnology textbook.

Let us reiterate some important points regarding protein structure from the Class XI Biotechnology textbook. Protein structure has been divided into four hierarchial levels to understand their organisation:

The linear order or sequence of covalently linked amino acid sequence is defined as **primary structure**. Depending on the nature and arrangement of the amino acids present different parts of the polypeptide chain form **secondary structures** like alpha helices and beta pleats. The **tertiary structure** organisation of these secondary structural elements occurs when these get compacted with each other to form compact spherical or globular units which are also thermodynamically stable conformations of these molecules in aqueous solutions (note cytoplasm is mainly water). In compaction several non-covalent interactions occur between the amino acid side chains. The **quarternary structure** is the association of two or more independent proteins/polypeptides via non-covalent forces to give a multimeric protein **(Fig. 2)**. The individual peptide units of this protein are referred to as subunits and they may be identical or different from one another.

The dominant forces which cause linear protein chains to undergo folding in space lies to a large extent in the chemistry of the amino acid residues they contain. Amino acids are broadly divided

into three main groups- **polar or hydrophilic** (eg. serine, glutamine), **charged** (eg. aspartate, arginine) and **hydrophobic** (eg. tryptophan, valine). Hence based on these features amino acid side chains can interact in space by a variety of non-covalent forces which is the basis of forming and stabilising protein structures in space. Let us examine some of the major non-covalent forces found in proteins.

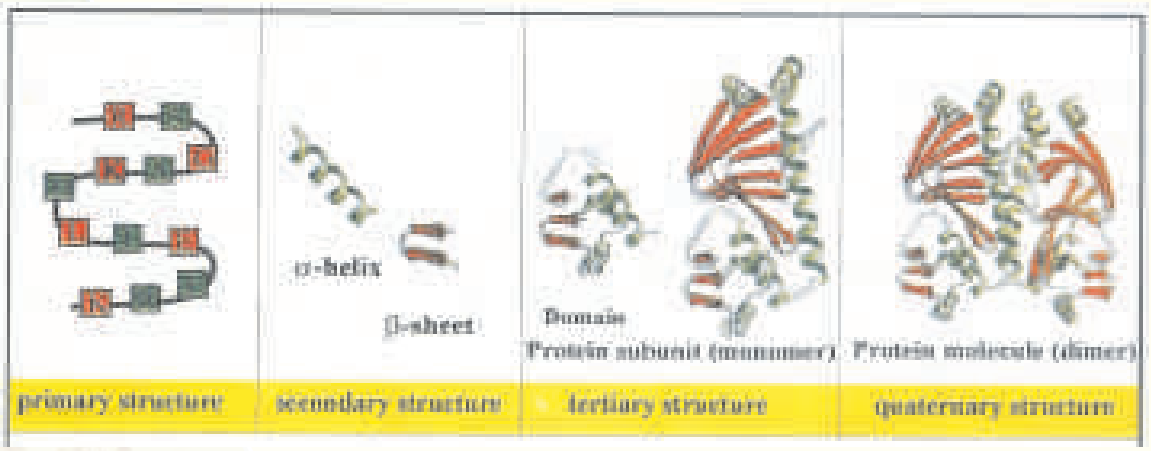


Fig. 2. Hierarchical organization in protein structure

Non-covalent bonds

The non-covalent interactions involved in organising the structure of protein molecules can be broadly divided into four categories:

- Ionic bonds
- Hydrogen bonds
- Van der Waals forces
- Hydrophobic interactions

Ionic bonds

These involve interactions between the oppositely charged groups of a molecule. For example the positively charged amino acid side chains of lysine and arginine can form salt bridges with the negatively charged side chains of aspartate and glutamate. These ionic interactions are also known as **salt bridges** because these are dominant bonds found in salts like sodium chloride wherein the positively charged sodium ion interacts with the negatively charged chloride ion. However, although ionic bonds have similar strengths to covalent bonds in vacuo, the bond strength of ionic bonds is vastly reduced in water due to the insulating qualities (dielectric strength) of water. Ionic bonds are highly sensitive to pH and salt concentration.

Hydrogen bonds

Hydrogen bonds are formed by "sharing" of a hydrogen atom between two electronegative atoms such as Nitrogen and Oxygen. In this case strongly polarised bonds between hydrogen and a small, very electronegative atom (N,O or F) allow a strong dipole-dipole bond to be formed with another small very electronegative element (N, O or F, **Fig. 3)**. Importantly, the very small sizes of these elements also allow them to approach each other so closely that a partial covalent bond is also formed (*e.g.*O-H---N). It is to be noted that the partial covalent character means that these bonds (H-bonds) are directional and strongest when the nuclei of all three involved atoms are in a linear arrangement.

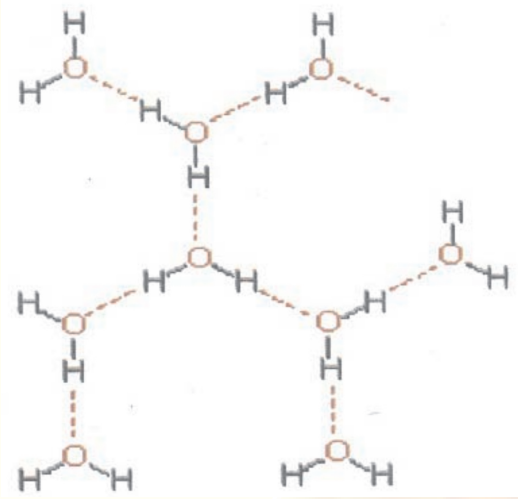


Fig. 3. Hydrogen bonding network in water

Van der Waals forces

These forces are weak attractions (or repulsions) which occur between atoms at close range. The Van der Waals types of forces are essentially contact forces, proportional to the surface areas in contact. These forces are of little significance at a distance due to the rapid 1/r6 (r is the inter-atomic distance) fall off. Even though weak, these bonds can be important in macromolecules because the large surface areas involved can result in reasonably large total forces.

Hydrophobic interactions

Hydrophobic interactions can be best explained by taking an example of oil in water. The oil tends to separate out fairly quickly, not because the oil molecules "want to get together", but because the water forces them out. The hydrophobic interaction is thus a manifestation of hydrogen bonding network in water. In water, each molecule is potentially bonded to four other molecules through H-bonds **(Fig. 3)**.

If a non-polar molecule, which cannot participate in hydrogen bonding, or in electrostatic interactions with water molecules, is added into water , a number of hydrogen bonds will be broken and not replaced. Since hydrogen bonds are favourable interactions, there will be an energy cost to putting non-polar molecules into water. Water therefore forces these molecules out of solution to minimise the surface of contact and thus the number of hydrogen bonds which are broken. Such forces known as hydrophobic forces are among the most important in driving proteins to fold into compact structures (globular) in water. Also, these forces are responsible to make different proteins assemble together to form structures found in muscles, membranes and

other organs. In proteins therefore, hydrophobic regions are preferentially located away from the surface of the molecule and form the interior core of the protein.

5.2.3. Structure-Function relationship in Proteins

As you have now learnt about how various forces drive proteins to assume characteristic shapes, it is worthwhile to consider why shape is paramount to the function of a protein. We will look at two proteins- an enzyme, chymotrypsin and the oxygen carrying protein, haemoglobin, to emphasise the importance of protein structure in its function.

Chymotrypsin, a proteolytic enzyme

As injested food makes way into the duodenum from the stomach, the proteins encounter a fierce proteolytic duo- trypsin and chymotrypsin which precisely cut the linear chains into short peptides which later on are acted upon by peptidases to release amino acids. Chymotrypsin, which hydrolyses peptide bonds following bulky aromatic amino acid residues in polypeptides is actually synthesised in the pancreas and through the pancreatic duct released into the duodenum. Have you wondered why this enzyme being a powerful proteolytic enzyme does not end up cutting cellular proteins within the pancreas itself? Nature has ensured that chymotrypsin and other proteolytic enzymes are synthesised as inactive harmless precursors known as zymogens which are then activated when required only in the duodenum, their site of activity, a process called *in-situ* activation. This activation in molecular terms results in an alteration in its shape so that it may now be able to interact with its substrate. The inactive precursor enzyme is termed chymotrypsinogen and the fully active enzyme is called chymotrypsin. The enzyme chymotrypsin is made up of a linear chain of 245 amino acids interrupted into three peptides- A,B,C. The protein folds into a globular structure. In the 3-D structure of the enzyme three important amino acid residues, his57, asp102 and ser195 come close together in space **(Fig. 4)** which allows a "charge relay system" to operate as indicated in **Fig 5**. The negatively charged asp102 is able to hydrogen bond with the adjacent his57 partially borrowing the hydrogen ion from the latter. The his57 makes good its partial hydrogen ion loss to aspartate by attracting a hydrogen ion from the adjacent ser195 through the his57 residue much like a relay race where the baton is passed from one member to another, the difference here being that the baton is a charge.

Normally the hydroxyl group of a serine residue is not acidic (pKa 12) and this is true for all other serine residues of chymotrypsin; only ser195 becomes acidic due to the unique constellation of the three amino acid residues because the protein has folded uniquely in space. You may be curious about the importance about an acidic serine residue. The negatively charged oxygen anion is able to make a nucleophilic attack on the carbonyl carbon of the peptide bond of its substrate, loosening it so that a water molecule can hydrolyse the bond **(Fig. 5)**. The specific site of chymotrypsin (recall that the enzyme is specific to aromatic residues) is a large space created

within the enzyme active site and lined by hydrophobic residues which therefore only allow bulky aromatic, hydrophobic amino acids to bind. This binding brings the susceptible peptide bond close to the attacking ser195 residue. In chymotrypsinogen, the substrate binding site is blocked and hence the enzyme is inactive. *In-situ* activation of trypsin involves a proteolytic cut in chymotrypsinogen which results in a conformational change, exposing the substrate binding pocket.

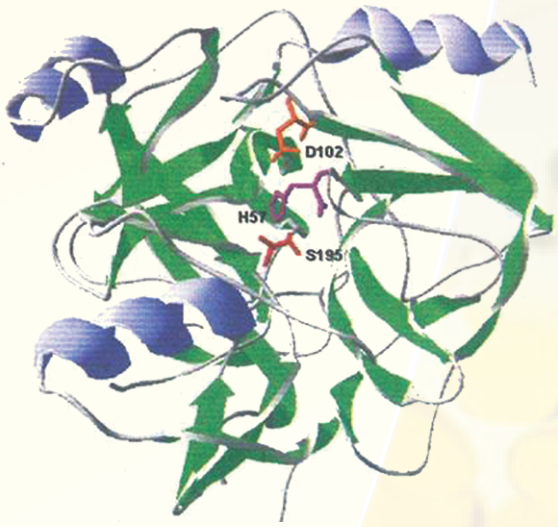


Fig. 4. Three dimensional structure of chymotrypsin

The interesting thing is that when nature has found a useful folding pattern which can cause hydrolysis of protein substrates, it repeats this in a variety of other enzymes. Trypsin, subtilisin (a proteolytic enzyme found in *B. subtilis*, a bacterium), thrombin (a proteolytic blood clotting factor) and the brain enzyme, acetyl choline esterase all have a reactive serine residue which is central to the catalytic mechanism.

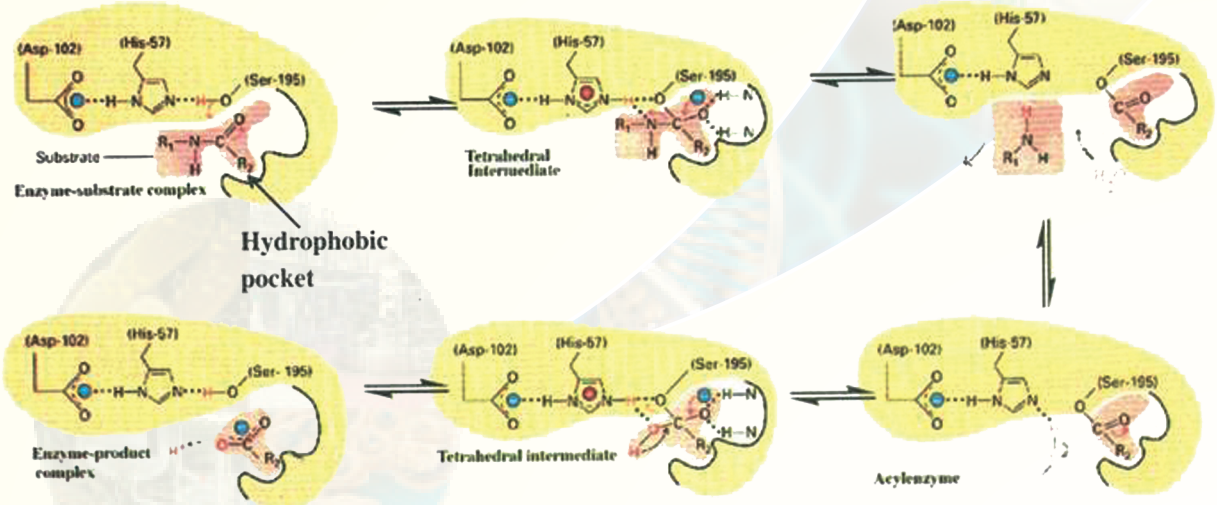


Fig. 5. Charge relay transfer in chymotrypsin. R_2 = aromatic amino acid; R_1 = any other amino acid.

Certain organophosphate compounds can selectively react with an acidic serine residue thereby knocking off enzyme activity. Nerve gas which was unfortunately used in the first world war had volatile serine alkylating compounds which inactivates the brain enzyme acetyl choline esterase

leading to death. Nowadays derivatives of organophosphates such as malathion and parathion which are not toxic to humans are used as mosquito repellants (Mortein, Good Knight) by effecting nerve transmission in insects.

Molecular Disease- Sickle cell anaemia

Sickle cell anaemia is a disease prevalent in parts of Africa and India where malaria is also endemic. The red cells of the patient have a pronounced morphological change and resemble the shape of a farmer's sickle and thus the name of the disease. Because these unusually shaped red cells have impaired oxygen carrying capacity and further get stuck in the small capillaries they lead to the anaemic conditions observed in patients. Interestingly such sickled RBCs resist malarial infection and hence offer some selection unfortunately for malaria to be co-prevalent with sickle cell anaemia. One of the first attempts to study the molecular basis of sickle cell anaemia was to compare the electrophoretic mobility of normal (Hb) and sickle cell haemoglobin (scHb). On finding that Hb moved faster than scHb, Linus Pauling predicted that the latter differed in a charged amino acid. This was confirmed by V. M. Ingram in 1957 who pioneered a useful technique called protein finger printing in the famous Laboratory of Molecular Biology (LMB) at Cambridge, UK. LMB has been the Mecca for protein sequencing, DNA sequencing, X-ray crystallography, deduction of the Double helix structure of DNA, Hybridoma technology and Nematode developmental studies. Established in 1952 under the leadership of Max Perutz (Received Nobel Prize for the structure of Haemoglobin) this institution has produced 9 Nobel Prize winners.

Protein Finger printing- Peptide Mapping

This technique involves the generation and 2-D analysis of peptides from a protein. Each protein has a unique peptide map (2-D analysis) and hence serves as a fingerprint for the protein. The steps involved in generating a peptide map/fingerprint are as follows **(Fig. 6)**.

1. Pure Hb and scHb are taken separately into test tubes.
2. The Hb and scHb are digested with the proteolytic enzyme trypsin which cleaves the protein after basic amino acid residues Arg and Lys.
3. Two separate strips of Whatman filter paper are spotted with Hb and scHb tryptic peptides and the peptides allowed to separate using the technique of paper electrophoresis at pH 2.0. Highly charged peptides will migrate more towards the anode/cathode.
4. The paper strips are dried, attached to larger squares of Whatman paper and chromatographed at right angles to the electrophoretic direction using a solvent system Butanol: Water:Acetic acid. In such a system peptides will separate based on their partition coefficient between the solvent and paper which is dependant on the relative hydrophobicity of the peptides. More hydrophobic peptides will move with the solvent to longer distances.

5. The chromatograms are dried and stained with a suitable visualisation reagent like Ninhydrin wherein peptide containing regions appear as orange yellow spots.

6. The peptide map for Hb and scHb are compared and it was found that one peptide was differently placed in the scHb map.

7. On examining this peptide and determining its amino acid sequence, Ingram found that it had a valine substitution for glutamic acid in the peptide.

   The single substitution of valine for glutamic acid (val/glu are at the 6th position of the haemoglobin beta chain) dramatically changes the structure of scHb making it form fibres within the RBC resulting in the deformation of the cell (sickling). Since the disease was due to a molecular alteration the term molecular disease was applied.

Peptide mapping became a useful technique to compare similar proteins from different sources. Slowly the information became too vast and computers were used to store this data into databases so that homology searches could be made. The protein fingerprinting data has been further augmented with new databases containing 2-D electrophoresis patterns of entire proteins from a given cell type, a technique developed by O'Farrel.

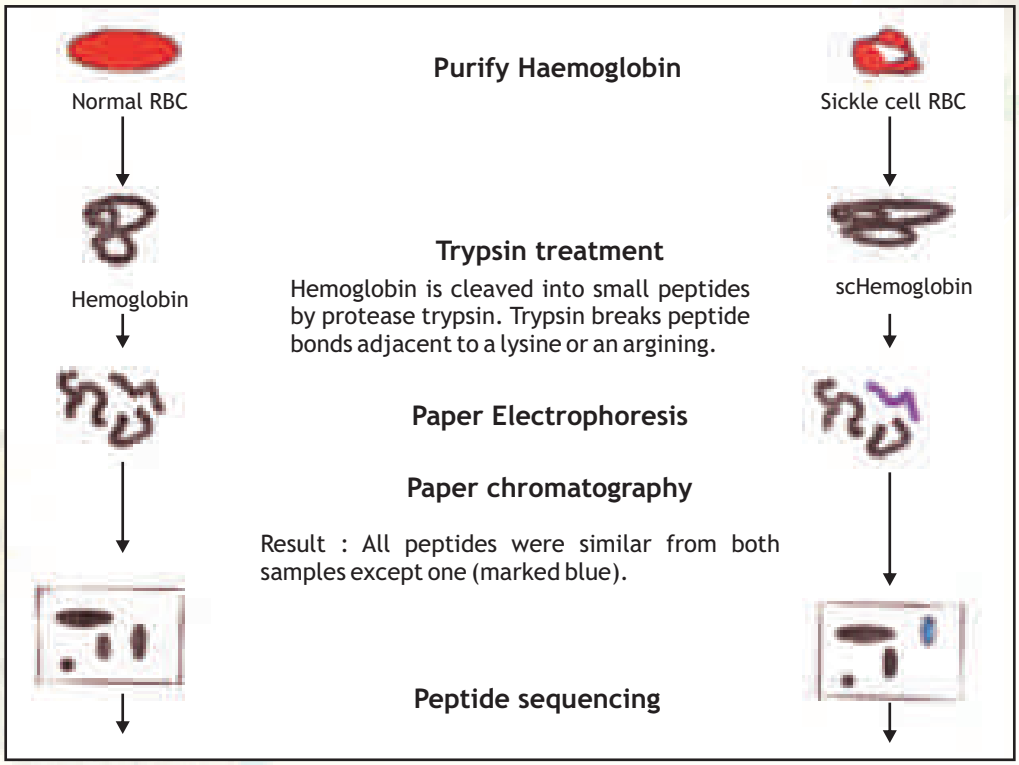


Fig. 6. Protein fingerprinting

2-D Gel Electrophoresis

Two different techniques are combined in this procedure- Isoelectro focussing (IEF) and SDS-PAGE

(Fig. 7).

In simple electrophoresis, the mobility of proteins is due to their charge, which is pH dependant. At its isoelectric pH (pI), a protein does not possess any charge and thus will not move in an applied electric field. This feature is exploited in the technique of IEF, which separates proteins on the basis of their different pI values. Usually IEF is performed in thin tube gels. A pH gradient is set up within the IEF gel by the inclusion of polymeric buffers known as ampholytes. These, like proteins have many positive and negative charges and hence varying pIs. Because of their smaller sizes they move rapidly in an electrophoretic run setting up pH gradients when they come to rest at specific distances from the anode/cathode when they have no net charge. A protein sample from a cell or any other source is then electrophoresed within these tubes wherein the different proteins separate and migrate to their pI zones. The tubes containing the separated proteins is then laid on a SDS-PAGE slab gel and electrophoresis continued at right angles to the IEF direction.

In SDS-PAGE proteins separate on the basis of their size and hence at the end of this electrophoretic run proteins are separated into 2-D patterns with high resolution as two properties of the proteins have been exploited in their separation- charge and size. Proteins in the gels are stained with silver stains or other highly sensitive dyes and can be scanned, and pictures stored into computer databases for analysis.

5.2.4. Purification of Proteins

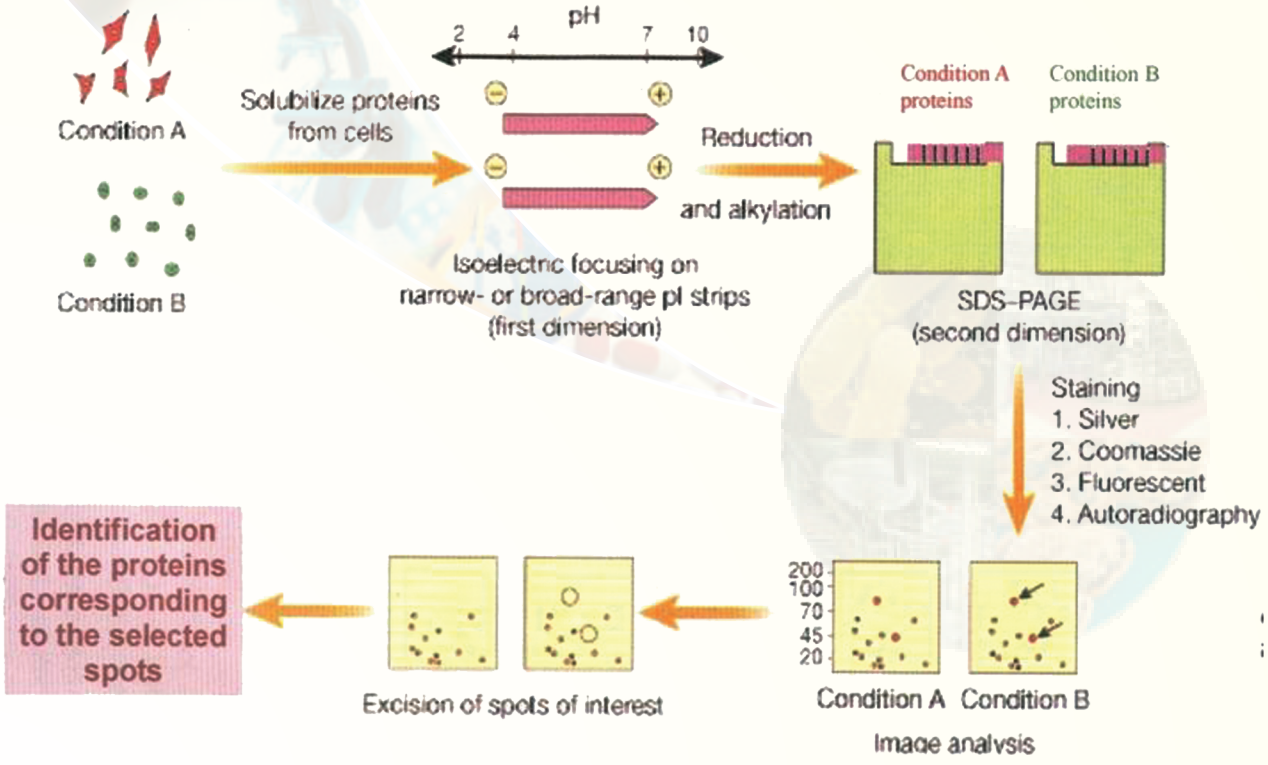


Fig. 7. Two dimentional gel electrophoresis

Isolation of a protein from a microbial culture, plant and animal sources involves various separation techniques (refer Textbook of Biotechnology for Class XI). These steps are collectively known as **downstream processing**. In spite of a large biodiversity of microbes we are restricted to certain bacteria/organisms which can be used as a source of protein as well as for introducing genes. These microorganisms are designated as "generally regarded as safe" (GRAS). **GRAS listed organisms are non-pathogenic, non-toxic and generally should not produce any antibiotics.** Similarily, plant tissue derived enzymes which have application in the food industry must be obtained from only non-toxic, edible plant species. One of the best known industrially useful enzymes is **papain** obtained from the latex of the green fruit and leaves of the Papaya tree. This enzyme finds application in meat tenderisation, clarification of beverages, digestive aids and wound cleaning solutions.

The existence of slaughter house facilities in which large number of animals are regularly processed to provide meat has also facilitated the collection of significant quantities of a particular tissue required as a protein source. Insulin is a classic example of a peptide hormone obtained from the pancreas of cows and pigs till the 1980's. Classical biotechnology required the slaughtering of over 100 pigs or 15 cows to meet the insulin requirements for one diabetic person for about one year. A rough estimate puts the number of diabetics in India by the year 2020 as 20 million! Obviously the requirement of insulin cannot be met by slaughtering pigs and cows. Fortunately, the advent of genetic engineering has ensured the availability of recombinant human insulin expressed in bacteria. Attempts are on to create transgenic animals by direct micro-injection of DNA into ova or stem cells and produce insulin and other proteins in milk on a commercial scale. This technology is called Molecular **Pharming** (Producing pharmaceuticals using genetically modified plants or animals). Advantages of producing recombinant proteins in milk are:

1. High production capacity.
2. Ease of source material collection (milking cows).
3. Moderate capital instrument requirements and low operational cost.
4. Ease of production including purification and scale-up.

Some medically useful peptides such as oxytocin have also been produced by direct chemical synthesis. In spite of different sources of proteins, the general principles of purification are similar. The overall approach and techniques are outlined in Unit-II, Chapter-III of the class XI, Textbook of Biotechnology. The exact details of the purification scheme for any given protein will depend upon a number of factors such as:

1. Exact source material chosen and location of the target protein (intracellular or extracellular).

2. Quantity of protein required and hence amount of raw material processed.
3. Physical, chemical and biological properties of the protein.

Calculation of amount of bacterial ferment required

Question: An *E. coli* cell produces at least 2000 different proteins. One of these is our enzyme of interest produced at a level of 3000 molecules per cell under optimum conditions. If we have to purify 1g of this intra-cellular enzyme, estimate how many cells of bacteria will be required theoretically? It is given that the molecular weight of the enzyme of interest is 1,00,000.

Answer: 1,00,000 g of the protein of interest corresponds to 1 mole of enzyme which corresponds to 6.023×10^{23} molecules (Avagadro no.)

Hence 1 g of enzyme has $1/1,00,000 \times 6.023 \times 10^{23}$ or 6.023×10^{18} molecules .

3000 molecules of the enzyme are present in one cell.

Therefore, 6.023×10^{18} molecules are present in $6.023 \times 10^{18}/3000 = 2.007 \times 10^{15}$ cells.

Question: Assuming that the bacterial cell is a cylinder (d = 1µm, h= 2µm) calculate (a) the total packed cell volume of *E.coli* required to produce 1 g of intra-cellular enzyme (b) the volume of the fermentor required if the maximum cell concentration inside the fermentor is 5% (cells need space to multiply).

Answer: Volume of a single bacterium = πr2h (cylinder volume) = $3.142 \times 0.5 \times 0.5 \times 10^{-12} \times 2 \times 10^{-6}$

(note 1µm = 10^{-6}m, r = ½d) $= 1.57 \times 10^{-18}$ m^3

2.007×10^{15} cells (see previous answer) would have a volume of $2.007 \times 10^{15} \times 1.57 \times 10^{-18}$ m^3

$= 3.15 \times 10^{-3}$ m^3 = 3.15 L Answer (a)

(1L = 10^{-3} m^3)

Answer (b) 100% concentration = 3.15 L

Therefore 5% concentration = 100/5 x 3.15 = 63 L

Volume of the fermentor required would be more than 63 L (30% extra space) about 82 L.

The source material chosen will dictate the range and type of contaminants present in the starting material. If the protein is extracellular, then one needs to separate the cellular components and process the medium to isolate the protein of interest. However if the protein is intracellular then the choice of method of cell disruption will depend on the cell type. Plant and fungal cells require harsher breakage methods; animal cells are easier to break because of no cell wall. Bacterial cells being very small require high pressure techniques. Once the proteins are

released into suitable buffered solutions a variety of physico-chemical techniques are applied to selectively purify the protein of interest from the others.

Genetically engineered proteins are often tagged with certain molecules in order to confer some very pronounced physico-chemical characteristics on the protein of interest. This renders its separation from contaminants more straightforward. The ability to detect and quantify the total protein levels is an essential pre-requisite to the purification and characterisation of any protein. A typical purification scheme can be analysed as follows **(Table 1)**.

Table 1. Typical purification table

Procedure	Total protein (mg)	Activity (units)	Specific activity units/mg
Crude extract	20,000	40,00,000	200
Precipitation (salt)	5,000	30,00,000	600
Precipitation (pH)	1,000	10,00,000	1000
Ion-exchange chromatography	200	8,00,000	4000
Affinity chromatography	50	7,50,000	15,000
Size exclusion chromatography	45	6,75,000	15,000

Note that the last column is a good indication of whether a purification step is useful or not. This is because as a protein is purified its specific activity increases because the denominator should ideally decrease as irrelevant proteins are removed and only the specific protein/enzyme of interest is concentrated. Hence from the given table it is apparent that the step following affinity chromatography, size exclusion chromatography, is redundant as the specific activity does not change. For activity measurements it is also important to choose a proper assay method reflecting the sensitivity required and further the method should be specific. In the case of proteins absorbance measurements at 280 nm is easy, fast and non-destructive procedure for monitoring the concentration.

Bioassays can sometimes be more sensitive than chemical assays. However one needs suitable standards of known bioactivity values to arrive at a correct activity of the unknown sample. Where a protein of biological interest is concerned, example insulin, bioassays are mandatory. If the sample particularily has to be injected other safety tests such as toxicity have to be performed.

Downstream Processing

After cells (bacterial, animal or plant) have grown to their requisite capacity in a fermentor it becomes necessary to harvest the cells or medium depending in which component the

recombinant protein is expressed and then purify the protein from other substances. These processes are part of downstream processing and because of the large amounts of source (a fermentor can be more than 1000 L capacity) bulk separation methods are used which are different from laboratory scale purification although principles involved are similar.

In the case of intracellular microbial proteins, cell harvesting is done by filtration or centrifugation from the fermentation medium, followed by re-suspension of cells in buffer or water with subsequent cell disruption. Most proteins obtained from plant and animal tissues are intracellular in nature. The initial step involves collection of the appropriate tissue, for example collection of blood to obtain proteins, collection of pituitary glands to obtain pituitary hormones etc.

Aqueous two-phase partition

When a crude cell homogenate is added to a biphasic mixture of dextran and polyethylene glycol (PEG) the cellular debris partitions to the lower, more polar and dense phase, dextran. Separation of the two phases achieves effective separation of cellular debris from soluble protein **(Fig. 8)**.

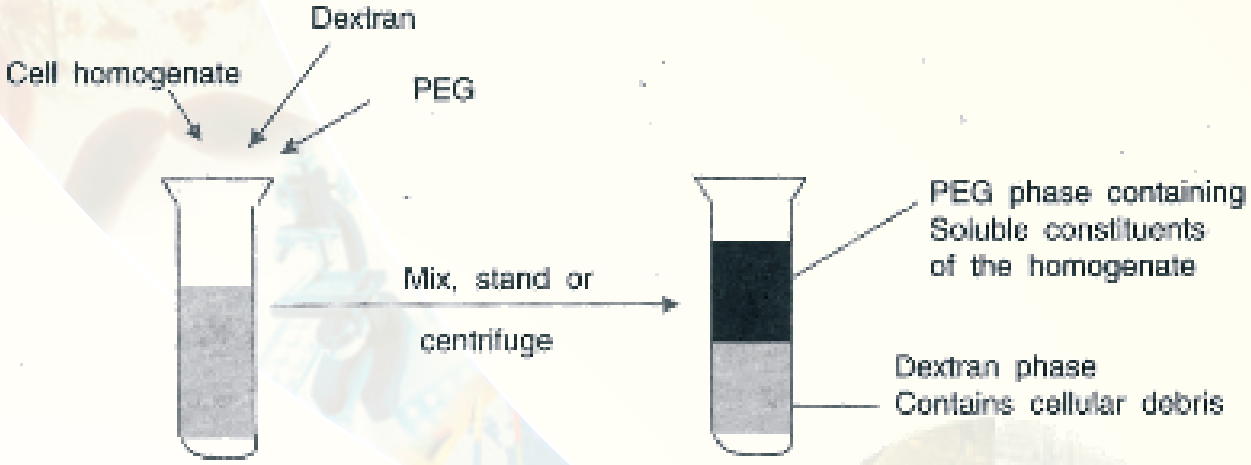


Fig. 8. Two phase separation

In some cases it is desirable and necessary to remove or destroy the lipids and nucleic acid of a cell homogenate as it may be a contaminant and can interfere with subsequent purification steps. The lipid layer can be removed by passage of solution through glass wool or cloth of very fine mesh size. Effective removal of nucleic acids may be achieved by precipitation or by treatment with nucleases.

For large scale application, concentration of extracts is normally achieved by precipitations, ion-exchange chromatography or ultrafiltration.

At any given pH value proteins display either a net positive, negative or no charge. Using these parameters different protein molecules can be separated from one another by judicious choice of pH, ionic strength and ion-exchange materials.

All efforts should be made to maximise protein stability during various steps. Some of the general conditions which may be followed are:

1. Maintenance of a specific pH value range of buffered solutions in which a protein is maximally stable.
2. Maintenance of physiological conditions (%CO_2 for animal cell culture and temperature).
3. Use of inhibitors to prevent the action of proteolytic enzymes.
4. Avoidance of agitation or addition of chemicals which may denature the target protein.
5. Minimise processing time.

Industrial scale production of proteins

The laboratory scale design cannot be scaled up to industrial scale directly. The following points need attention for industrial scale production:

1. Bulk purchase of chemicals and other raw materials would bring down costs.
2. The labour cost decreases sharply with increase in production.

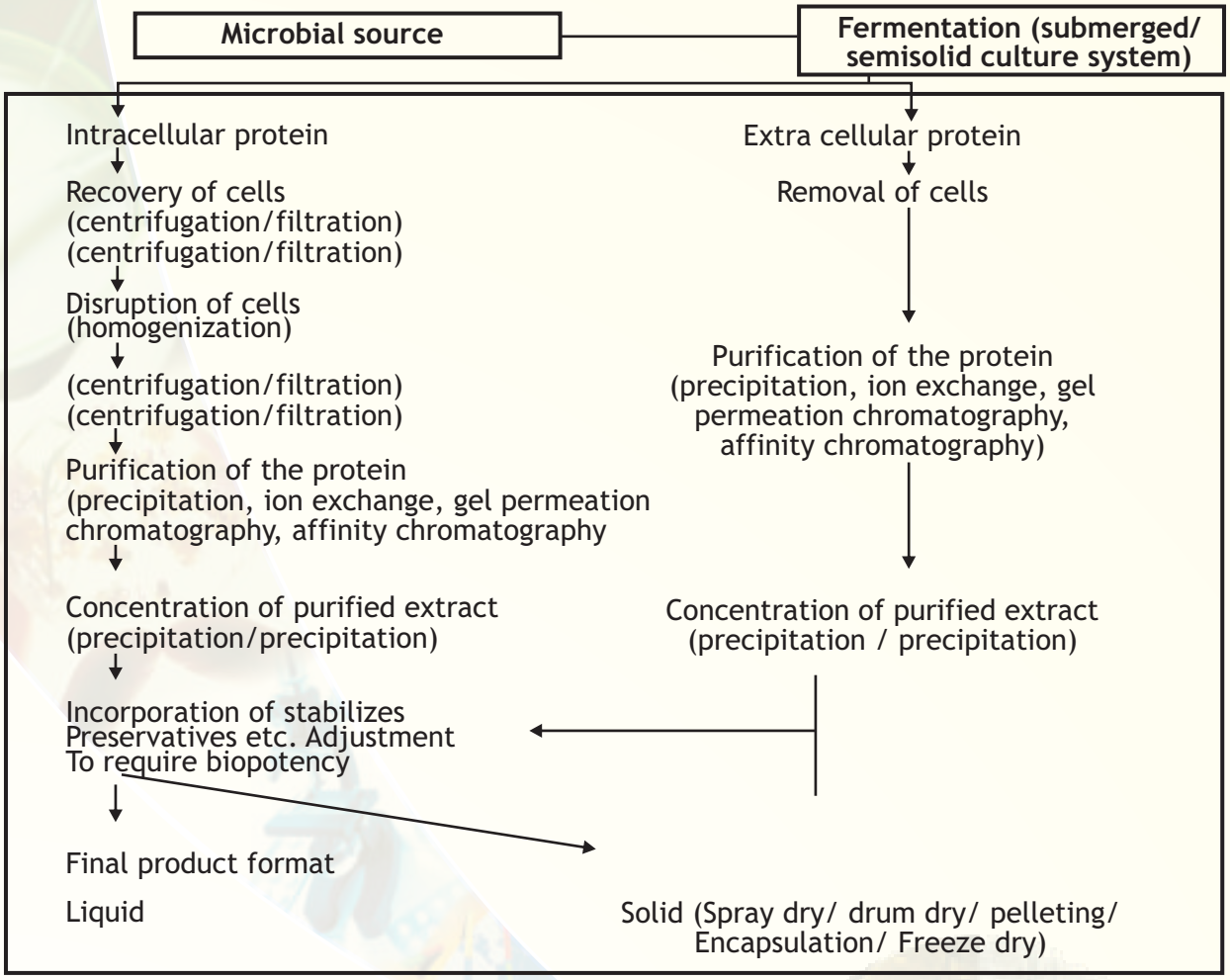
Most large scale process equipment such as holding vessels and transfer pumps are constructed from stainless steel or plastics, such as polypropylene. Glass vessels, so commonly used in laboratory scale culturing techniques are seldom used for large scale preparatory work. Materials used for large scale culturing must be inert and resistant to the corrosive action of any chemical used during the process. They should not allow any leaching of potentially toxic metals or chemicals into the product stream. It is useful to remember that any commercial plant has to have good GMP (good manufacturing practice) . Any downstream processing requires the approval of a regulatory authority in the form of a license to produce and market proteins designed for use in the food or health care industry. This ensures that the processing procedures are based upon established, validated methodologies. Generally 80% of the overall production cost are due to steps in downstream processing and quality assurance.

A generalised downstream processing scheme used in the production of bulk protein/enzyme from microbial sources is given in **Fig. 9.** Similar steps can be applied for animal and plant sources.

Special techniques for therapeutic /diagnostic proteins

These proteins must be purified to a very high degree especially for use in parenteral (injectable) administration. They also have to be sterile products (free of bacterial and fungal contamination) and hence can be administered by injection, infusion or implantation.

Fig. 9. A typical flow sheet from source to product.



5.2.5. Characterisation of Proteins

Techniques listed below characterise proteins with respect to properties such as mass, isoelectric charge, amino-acid sequence etc. Alongside they can also detect impurities as these are very sensitive techniques requiring small amounts, often micrograms of the sample. The first four techniques have already been discussed in this chapter.

1. Electrophoretic techniques, SDS/PAGE.
2. Fingerprinting.
3. Two dimensional gel electrophoresis.
4. Protein sequencing.
5. Mass spectrometry.

Mass spectrometry

Mass spectrometry (MS) has emerged as an important tool in biotechnology. It is extremely useful in obtaining protein structural information such as peptide mass or amino acid sequences. It is also useful in identifying the type and location of amino acid modification within proteins. One of the major attractions of mass spectrometry is that as little as picomoles (10^{-12}) of a protein sample can be analysed. A mass spectrometer is an analytical device that determines the molecular weight of chemical compounds by separating molecular ions according to their mass/charge ratio (m/z) ratios. The molecular ions are generated either by a loss or gain of a charge (e.g. electron ejection, protonation or deprotonation). After the ions are formed they can be separated according to their m/z ratio and finally detected. The process of ionisation, ion separation and detection in a mass spectrum can provide molecular weight or even structural information. A sample M with a molecular weight greater than 1200 D can give rise to multiple charged ions such as (M+nH)n+. Proteins/peptides have many suitable sites for protonation as all the backbone amide nitrogen atoms could be protonated theoretically as well as certain amino acid side chains such as lysine and arginine which contain primary amine functional groups.

A schematic diagram of the various parts of a mass spectrometer is indicated in **Fig.10.** Basically a vapourised sample of a protein or peptide is introduced into the instrument wherein it undergoes ionisation. The charged molecules are then electrostatically propelled into a mass analyser (filter) which separates the ions according to their m/z ratio. The signal received upon detection of the ions at the detector is transferred to a computer which stores and processes the information.

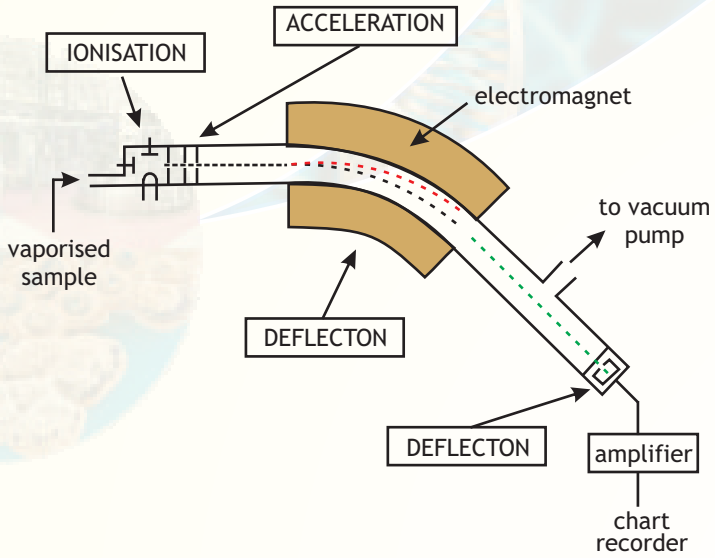


Fig. 10. An outline of a mass spectrometer.

The goal of mass spectrometric analysis of biomolecules like peptides and proteins is to create gas phase ions from polar charged molecules which are generally non-volatile. A popular method called Matrix Assisted Laser Desorption Ionisation (MALDI) is used to volatalise and protonate peptides and proteins. In this procedure, the sample is transferred from a condensed phase to a gas phase with the help of a solid matrix. Ion formation in MALDI is achieved by directing a pulsed laser beam onto a sample suspended or dissolved in a matrix. The matrix plays a key role in this technique by absorbing the laser light energy and causing the matrix material to vaporise. In the gas phase, the matrix plays a role in sample ionisation. The charged molecules are directed by electrostatic lenses from the ionisation source to the mass analyzer.

Multiply charged ions on analysis show patterns as indicated in Fig. from which molecular mass can be deduced as also indicated in the legend below **Fig. 11.**

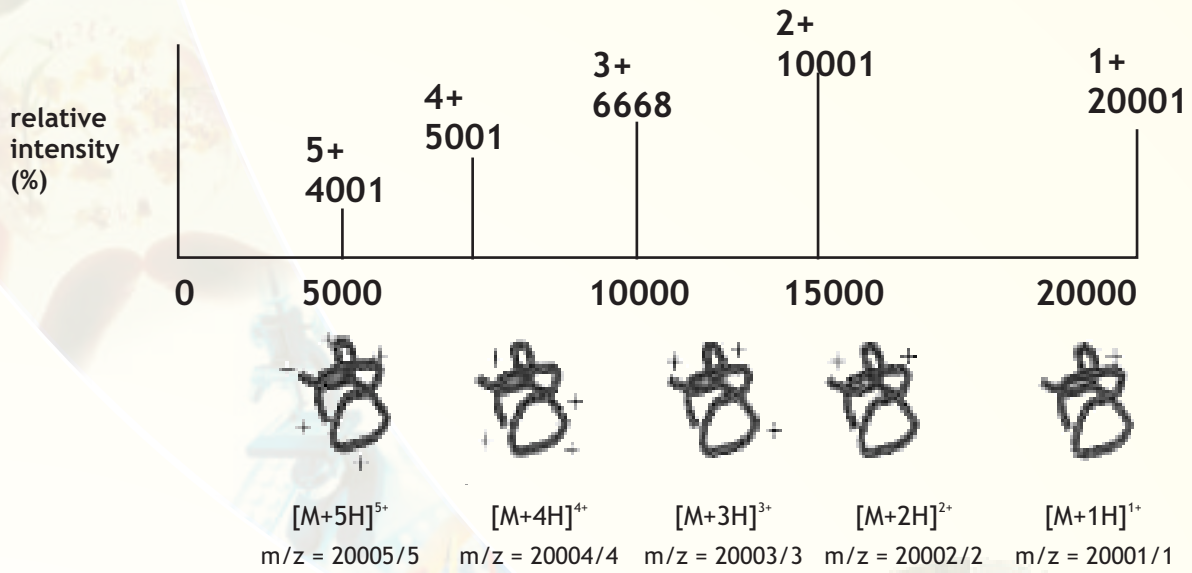


Fig.11. The same protein with a molecular weight of 10,000 contains 5, 4, 3, 2 and 1 charges. The mass spectrometer detects the protein ions at m/z = 2001, 2501, 3334, 5001 and 10,001 respectively.

Typically for protein identification, a crude extract is separated on a 2D gel. After visualising the proteins on the gel, protein spots are excised and used for mass analysis by the MS technique described above. For doing so the protein in the 2D gel after extraction from gel is either used intact or it is cleaved into small peptides with a protease like trypsin which makes the mass analysis easier. These peptides are separated on an on-line liquid chromatography system before introduction into the mass spectrometer. Liquid chromatography techniques like ion exchange, affinity or reverse phase column chromatography can be used to separate the peptides. These peptides are either sequenced directly or the mass of peptides is analysed using database searches (see Bioinformatics unit). With the human genome sequencing project, it is now possible to identify new proteins by combining the mass spectrometric information with the genomic

information using Bioinformatics tools.

5.2.6. Protein Based Products

From the commercial point of view, proteins may be classified into the following categories.

1. Blood products and vaccines.
2. Therapeutic antibodies and enzymes.
3. Therapeutic hormones and growth factors.
4. Regulatory factors.
5. Analytical application.
6. Industrial enzymes.
7. Functional non-catalytic proteins.
8. Nutraceutical proteins.

Blood products and vaccines

Blood carries out several functions and is one of the best mediums for transportation in an animal. A better understanding of haematapoiesis (formation of blood cells) as well as factors responsible for blood coagulation has led to the discovery of several useful proteins. Several proteins from blood and plasma have been commercially available for decades. While these products have traditionally been obtained from blood donated by human volunteers, some are now produced by recombinant DNA technology. For example Factor VIII for treatment of Haemophilia A, Factor IX for treatment of Haemophlia B, Hepatitis B vaccine for prevention of hepatitis etc.

Therapeutic antibodies and enzymes.

Polyclonal antibodies have been used for more than a century for therapeutic purposes. More recently monoclonal antibody preparations as well as antibody fragments produced by recombinant DNA technology have found therapeutic use. For example tissue plasminogen activator (t-PA) is a proteolytic enzyme used to digest blocks in arteries following myocardial infarctions. A monoclonal antibody OKT-3 is used to prevent rejection following kidney transplantation because the antibody blocks those immune cells which attack foreign grafts.

Therapeutic hormones and growth factors

A number of hormone preparations have been used clinically for many decades. Though insulin was prepared from the pancreas of cows and pigs, the ability to genetically transfer human insulin gene into bacteria and the ability to modify amino acid residues (protein engineering) has

facilitated the development of modified forms which are faster acting like humulin. Humulin acts in 15 min unlike pig insulin which takes 3 hours. Another growth factor- platelet derived growth factor has been approved for diabetics who develop skin ulcers. Several other growth factors are under various stages of clinical trials.

Regulatory factors

Several new regulatory factors were discovered that did not fit the classical definition of a hormone. Initially they were known as cytokines. These include interferons, interleukins, tumor necrosis factor and colony stimulating factors. The interferon family of INF alpha, beta and gamma have found widespread therapeutic application; interferon alpha is used for treatment of Hepatitis C, beta for Multiple Sclerosis and gamma for Chronic Granulomatous disease.

Analytical applications

Enzymes and antibodies have found a range of analytical applications in the diagnosis of diseases; hexokinase for quantitative estimation of glucose in serum, uricase for uric acid levels in serum, horse radish peroxidise and alkaline phosphates in ELISA etc.

Industrial enzymes

Proteolytic enzymes constitute an 8000 crore annual market for industrial enzymes. They find application in the beverage industry, detergent industry, bread and confectionary industry, cheese production, leather processing and meat industry. Alcalase is an enzyme used in the soap industry, papain is used in the beverage industry, glucose isomerise in the confectionary industry and chymosin is used in the cheese industry.

Functional non-catalytic proteins

Functional non-catalytic proteins are those which have properties such as emulsification, gelation, water binding, whipping and foaming etc. **(Table 2)**. For example kappa casein, a component of casein is involved in micelle stabilisation of milk proteins and keep the proteins suspended uniformly in milk because it behaves like a lipid molecule (2/3rd of the protein is hydrophobic). The food industry has exploited these non-catalytic proteins as illustrated in the below.

Table for whey protein.

TABLE : 2

Functional Property	Mode of action	Food System
Whipping/Foaming	Forms stable film	Egg less cakes, desserts, whipped topping
Emulsification	Formation and stabilization of fat emulsions	Vegetarian sausages, salad dressings, coffee whiteners, soups, cakes, infant food formulas, biscuits.
Gelation	Protein matrix formation and setting	Meat, baked goods, cheeses
Viscosity	Thickening, water binding	Soups, gravies, salad dressings
Water binding	Hydrogen bonding of water; entrapment of water	Meats, sausages, cakes, breads
Solubility	Protein solvation	Beverages
Browning	Undergoes Maillard reaction (on heating, the amino groups of protein react with aldehyde groups of sugars)	Breads, biscuits, confections, sauces
Flavour/Aroma	Lactose reacts with milk proteins	Baked goods, biscuits, confectionaries, sauces, soups, dairy products.

Nutraceutical Proteins

Nutraceutical is a word coined from combination of nutrition and pharmaceuticals. It has been observed that several nutritional proteins also have therapeutic functions. For example whey protein concentrates, lactose free milk (for lactose intolerant babies) and infant food formulations.

Where does one get the raw building materials such as amino acids needed to make all body proteins? During infancy we depend on milk. Baby milk formulations are also there (Amul, Lactogen etc.) which have been formulated to have similar composition as mother's milk. All these food materials provide the essential components nutritionally for growth and development during the first few months of our existence. A typical composition of milk from buffalo, human and cow sources is given in the **Table 3** from which baby milk formulations can be made to suit an infant.

Table 3. Composition of milk from buffalo, human and cow.

Constituents (per 100 ml of milk)		buffalo	human	cow
1.	Protein (g)	3.8	1.2	3.3
2.	Casein (g)	3.0	1.4	2.8
3.	Lactalbumin (g)	0.4	0.3	0.4
4.	Lactoglobulin (g)	0.2	0.2	0.2
5.	Fat (g)	7.5	3.8	3.7
6.	Lactose (g)	4.4	7.0	4.8
7.	Calorific value (K Cal)	100.0	71.0	69.0
8.	Calcium (mg)	203.0	33.0	125.0
9.	Phosphorous (mg)	130.0	15.0	96.0
10.	Chloride (mg)	112.0	43.0	103.0

From the **Table 3** it can be observed that milk contains several proteins, carbohydrates, lipids, vitamins, antibodies, minerals etc. It is interesting to note that human milk has nearly half the amount of casein as compared to cow and buffalo. Besides use of milk as a nutritional source, claims have been made to the effect that curd is beneficial in the management of some types of intestinal infections according to our ancient Sanskrit scriptures dating back to 6000 BC. Since time immemorial whey (liquid part of curds) has been administered to the sick for the treatment of numerous ailments. In 1603, Baricelli reported on the therapeutic use of cow's or goat's whey, sometimes mixed with honey and herbs. The spectrum of illnesses treated with whey include jaundice, infected skin lesions, genitor-urinary tract infections. Gallen and Hippocrates insisted on a minimum daily drinking of one litre of whey. Using modern scientific research it has been possible to explain these observations. Whey proteins result in the elevation of a tripeptide glutathione(gamma-glutamyl cysteinyl glycine) in cells. This peptide is a reducing compound and has a broad range of functions including detoxification of xenobiotics and protection of cellular components from the effect of oxygen intermediates and free radicals. More recently curd has also been used as a pro-biotic (administered with antibiotics) because it is a good source of beneficial bacteria which can colonise the intestinal tract. **Table 4** gives the useful components of whey.

Table 4. Whey components.

α - Lactalbumin	Lactose
α - Lactoglobulin	Sialic Acid
Bovine Serum Albumin	Lactic Acid
Immunoglobulins	
Lactoferrin	Sodium
Lactoperoxidase	Potassium
	Calcium
	Magnesium
Protease peptones/polypeptides	Chloride
Free amino acids	Phospphate
Urea	Sulfate
	Citrate
Glycomacropeptides	Heavy Metals
Growth Factors	Milk Fat
	Globule
	Free Fat
	Lipoproteins

5.2.7. Designing Proteins (Protein Engineering)

Considerable interest exists in the biotechnology industry for the engineering of proteins with increased stability when exposed to harsh conditions like elevated temperature, organic solvents and reactive chemicals, often encountered in the industrial processes. Besides, it is of great interest to explore biological adaptations to environmental stresses such as high salinity, drought, cold etc. Therefore, in order to stabilise your favourite protein, it is essential to know the cause of inactivation.

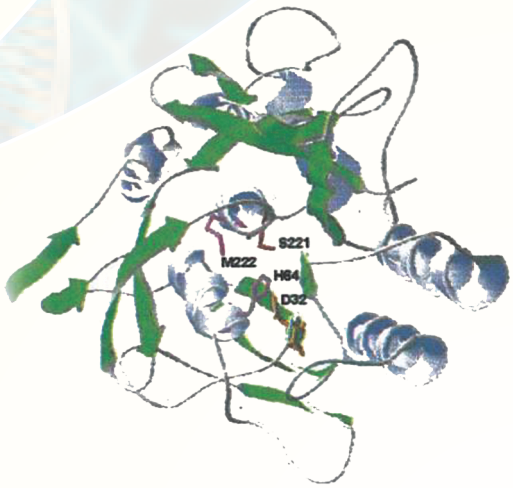


Figure 12. Subtilisin with the catalytic triad

Stability in a folded protein is a balance between the stabilising (mainly hydrophobic) interactions and the tendency towards destabilisation caused by the loss of conformational entropy as the protein adopts the unfolded form. The stability of a protein may however be changed by substituting amino acids that either favour stabilising interactions in a folded protein or destabilising interactions in an inactive protein. Numerous attempts have been made on different

proteins and enzymes in order to improve their properties for thermal and pH stability, solvent tolerance and solubility, catalytic potency etc. Given below is an example which has been successfully used in the detergent industry.

Improving laundry detergent Subtilisin

Subtilisin (27 kD) is a protease produced by bacteria that can digest a broad range of proteins that commonly soil clothing, see **Fig. 12**. The enzymatic activity of subtilisin is contributed by a catalytic triad, i.e., Ser221, His64 and Asp32 similar to chymotrypsin. Replacement of all three residues with alanine either singly or in combination results in significant loss of activity. Subtilisin represents the largest industrial market for any enzyme. To improve the efficiency of laundry detergents, detergent manufacturers supplement subtilisin in their products with various catchy slogans on the detergent box such as "stain cutter" or "biologically active enzymes".

The native enzyme subtilisin is easily inactivated by bleach (up to 90%). Careful studies showed that this inactivation was due to oxidation of the amino acid residue Methionine222 in the protein molecule **(Fig.12)**. Using site-directed mutagenesis of the subtilisin gene in *E.coli*, this methionine was substituted by a variety of other amino acids and the enzyme activity measured in the presence of bleach **(Table 5)**. It was observed that substitution of Met222 with Ala222 was the best in terms of activity and stability. Nowadays, many laundry detergents contain cloned, genetically engineered or recombinant subtilisin.

Table 5. Site-directed mutagenesis at codon position 222.

Codon-222	% activity w.r.t wild type	Codon-222	% activity w.r.t. wild type
Cys	138.0	Gln	7.2
Met	100.0	Phe	4.9
Ala	53.0	Trp	4.8
Ser	35.0	Asp	4.1
Gly	30.0	Tyr	4.0
Thr	28.0	His	4.0
Asn	15.0	Glu	3.6
Pro	13.0	Lie	2.2
Leu	12.0	Arg	0.5
Val	9.3	Lys	0.3

Creation of Novel Proteins

Conventional vaccines have utilised heat inactivated bacteria/viruses or their surface proteins to generate immunity against various specific diseases. Often it has been observed that some of the components of the vaccines have undesirable effects such as fever and rarely one has also heard of some of the components like virus actually causing the disease due to incomplete inactivation. Since proteins are the main molecules which provide the stimulus for immunity, attempts have been made to engineer proteins having minimum deleterious effects. The specific sequences of amino acids in the protein which stimulate immune response are known as epitopes. A recombinant vaccine based on selected epitopes may provide optimal design, scope for micromanipulation, unhindered supply and safety needed for an effective vaccine. Working on these lines, a novel synthetic gene has been assembled as a first step towards developing a subunit vaccine against Hepatitis B virus.

Improving nutritional value of cereals and legumes

The cereal grains and seeds of legumes constitute a major chunk of dietary protein requirement. The seed storage proteins are synthesised and accumulated throughout seed development to serve as source of amino acid reserves at the time of seed germination. High levels of such proteins in seeds would provide an enriched amino acid source for human consumption. However deficiencies in seeds of certain essential amino acids render the cereal grains or legumes unsuitable for a balanced diet. Supplementation of diet with essential amino acids from other sources therefore becomes essential. **Fig.13** gives the essential amino acid content of various cereals and commonly eaten food proteins. Essential amino acids are those which have to be obtained from food and cannot be made in our cells. From the data in the figure it is apparent that whey protein is superior to other sources especially with regard to branched amino acids- ile, leu, val, lys and trp. The branched chain amino acids (BCAA) are essential for the biosynthesis of muscle proteins. They help in increasing the bio-availability of high complex carbohydrates intake and are absorbed by muscle cells for anabolic muscle building activity. One of the theories is that during exercise the BCAAs are released from the skeletal muscle; the carbon skeleton part is used as fuel and the nitrogen part is used to make alanine which then goes to the liver where it is turned into glucose for energy. So for athletes who want to protect their existing mass, the idea is to take BCAA enriched foods before and after excercise. BCAAs reduce muscle breakdown and act as an energy source before and after exercise. Hence while maintaining exercise performance and delaying exhaustion BCAAs are very important for muscle growth. Nowadays an entire new area of sports medicine and nutrition prepare and recommend special nutrient drinks etc. which incorporate these principles. In the unit on plant tissue culture you will read how plant cereals have been genetically engineered for higher nutrient value in terms of proteins, vitamins etc.

Biological value (BV) measures the amount of protein nitrogen that is retained by the body from a given amount of protein nitrogen that has been consumed. It has been observed that the BV of whey proteins is the highest compared to rice, wheat, soya and egg proteins. Another index of protein value is the **protein efficiency ratio** (PER). PER is used as a measure of growth expressed in terms of weight gain of an adult by consuming 1g of food protein. The PER value of the following proteins are arranged in decreasing order- whey, milk, casein, soya, rice, wheat. The modern day approach for overcoming the nutritional deficiencies of seeds would be to engineer genes that would encode storage proteins with more of the nutritionally desirable amino acids either by inserting additional amino acids or substituting existing amino acids with new ones. Attempts are already being made on zein storage protein genes of maize to enhance its nutritional value. Introduction of entirely novel proteins that are highly enriched in specific amino acids is also being considered.

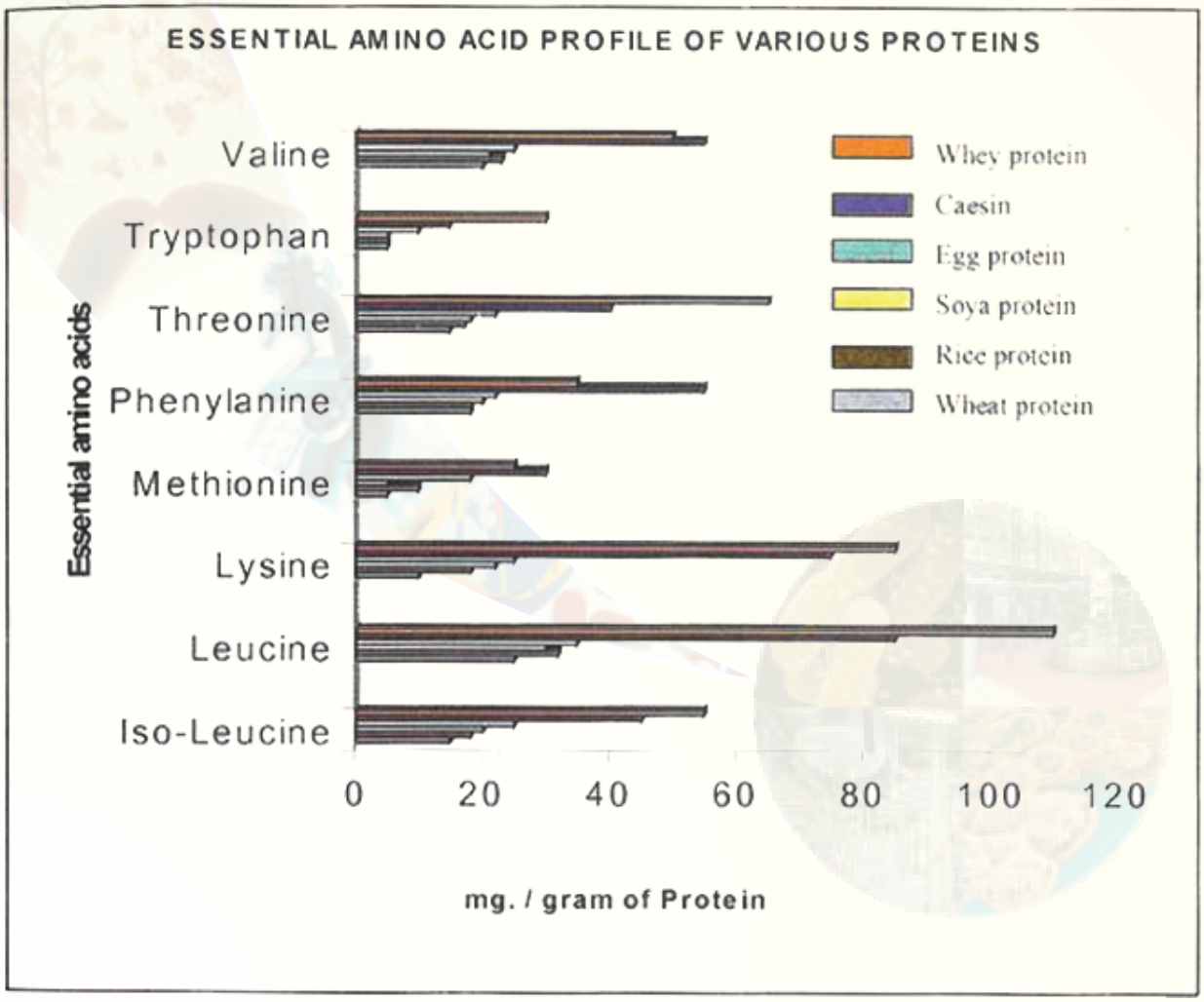


Fig. 13. The essential amino acids profile.

Review Questions

1. Name two human diseases caused by the absence of a protein.
2. What is the consequence if a protein is incorrectly folded? Give an example to illustrate your answer.
3. Distinguish between chymotrypsinogen and chymotrypsin.
4. Briefly explain how the serine residue in some enzymes can become acidic (reactive). Also suggest how you can confirm that a serine residue is involved in the catalysis.
5. Why is sickle cell anaemia called "Molecular disease"? How can sickle cell haemoglobin be identified?
6. What are the principles behind Isoelectric Focusing and SDS/PAGE techniques? Why is 2-D electrophoresis better than single dimension electrophoresis?
7. Define subunit, domain, and quaternary structure in proteins.
8. With an example explain the development of one protein based product.
9. What are non-catalytic functional proteins, therapeutic proteins and nutraceutical proteins? Give one example each.
10. Briefly discuss the use of designing a protein for any product.
11. What is the principle of MALDI-TOF? What is its main use in protein studies?
12. *E. coli* is a rod shaped bacteria about 2 µm long and 1 µm in diameter. The average density of a cell is 1 .28 g/ ml. Approximately 13.5% of the wet weight of E. coli is soluble protein. Estimate the number of molecules of a particular enzyme per cell if the enzyme has a molecular weight of 100,000 and represents 0.1 % of the total soluble protein. (Answer: 1626 molecules per cell).

References

1. Protein: Biochemistry and Biotechnology by Gary Walsh (2002), John Wiley & Sons, Ltd.
2. Practical Biochemistry: Principles and Techniques Edited by Keith Wilson and John Walker (2000), Cambridge University Press.
3. Biochemistry by Lubert Stryer, J.M, Berg and J.L. Tymoczko (2002). W.H. Freeman.
4. Principles of Biochemistry by Albert L. Lehninger, D.L. Nelson and M.M. Cox (2000), W.H. Freeman.
5. Biochemical calculations by Irwin H. Segel (1985), John Wiley & Sons, Ltd.
6. Proteins: Structures and Molecular Properties by Thomas E. Creighton (1992), W.H. Freeman.