In chapter 2, a brief introduction to proteins has been provided. Proteins are built up of amino acids, which form their building blocks. In all, there are 20 alpha amino acids found in proteins, and the specificity of a protein molecule is determined by their number and sequence. In different permutation and combinations of amino acids, about $10^7$ different types of protein molecules may exist in organisms.

The study of proteins falls under a specialized field called ‘proteomics’. It would be necessary to briefly discuss about this field and its importance.

10.1 PROTEOMICS

The term ‘proteomics’ stands for the systematic analysis of the entire protein complement produced by a genome, or by a cell or tissue type. Another term ‘proteome’ refers to all proteins produced by a species, much as the genome is the entire set of genes. Unlike the genome, the proteome varies with time and is defined as the ‘proteins present in one sample (tissue, organism, cell culture) at a certain point of time’.

Proteomics is a much bigger field than genomics, since it deals with one genome per organism, whereas several proteomes may exist in an organism. Besides, genome is a constant feature of an organism, while the proteome varies with nature of tissue, state of development, health or disease, and effect of drug treatment.

Gene-Protein Relationship

There is an increasing interest in proteomic technologies because DNA sequence information provides a static snapshot of various ways in which a cell might produce or use its proteins. With the help of Human Genome Project (HGP), around 30,000 genes have been identified in the human genome, but how come with so few genes an infinite variety of proteins can be produced. The possible solution is post-translational modification that occurs in the body. Once a protein is translated, it undergoes further processing, hence the number of proteins is much more than the number of genes (Fig. 10.1). Each protein can be identified by its unique amino acid sequence.
Proteins perform a variety of functions in the cell and proteomics enables correlations to be drawn between the range of proteins produced by a cell or tissue, and the initiation or progression of a disease state. Proteome research helps the discovery of new protein markers for diagnostic purposes and of novel molecular targets for drug action. Proteomics at a broad level attempts to catalogue and characterize these proteins, compare variations in their expression levels in various physiological states, study their interactions, and identify their functional roles.

10.2 PROTEIN STRUCTURE

A protein is a linear sequence of amino acids linked together by peptide bonds. The peptide is a chemical, covalent bond formed between the alpha amino group of one amino acid and the alpha carboxyl group of another amino acid. When two amino acids are joined by a peptide bond, a dipeptide is formed. The peptide bond has a partial double bond character and always in a \textit{trans}-configuration. Addition of more amino acids to the dipeptide extends the chain, forming oligopeptide or a polypeptide. Proteins do not exist as linear chains in native state, hence the linear sequence of amino acids folds up to form specific shapes (conformations) with three-dimensional arrangement of atoms, which can be determined by the amino acid sequence. Thus, four levels of protein structure have been recognized.

Chemical Bonding in Proteins

In the linear sequence of amino acids, the proteins are said to possess \textit{primary structure}, where amino acids are joined together by peptide bonds. The other covalent bonds in the primary structure are \textit{disulphide} bonds between cysteine residues and used for cross-linking purposes.

In secondary quaternary and tertiary conformations of proteins, non-covalent interactions are involved, which are identified in four categories.

1. \textit{Ionic bonds}: At any given pH, proteins have charged groups that may participate in binding them to each other or to other types of molecules. For example, negatively charged COO\(^-\) groups on aspartic acid (Asp) and glutamic acid (Glu) residues may be attracted by the positively charged free amino groups on lysine (Lys) and arginine (Arg) residues. Ionic bonds are highly sensitive to pH changes and salt concentrations.
2. **Hydrogen bonds**: Hydrogen bonds are formed when a strong electronegative atom (e.g., oxygen, nitrogen) approaches a hydrogen atom that is covalently attached to a second strongly electronegative atom. These are weak bonds.

3. **Van der Waals forces**: Van der Waal’s forces are very weak.

4. **Hydrophobic interactions**: The side chains (R groups) of such amino acids as phenylalanine and leucine are nonpolar, and hence interact poorly with polar molecules like water. For this reason, most of the nonpolar residues in globular proteins are directed toward the interior of the molecule, whereas polar groups of aspartic acid and lysine are projected on the surface exposed to the solvent. When nonpolar residues are exposed at the surface of two different molecules, it is energetically more favourable for their two ‘oily’ nonpolar surface to approach each other closely displacing the polar water molecules between them. The strength of hydrophobic interactions is not appreciably affected by changes in pH or in salt concentration.

### 10.3 DETERMINATION OF PROTEIN STRUCTURE

Proteins are highly complex molecules and their structure involves four levels of organization, as discussed in chapter 2. In order to understand the **primary structure** of proteins, it is necessary to characterize proteins in terms of molecular mass and constituent amino acids. Primary structure determination is essential for arriving at amino acid sequence in a protein. The steps that follow,

```
Source Material
    | Homogenization & Extraction
    |   | Precipitation
    |   |   | Chromatography
    |   |   | Concentration
    |   |   | Assessment
      Activity
      Purity
```

*Fig. 10.2 Steps in protein purification*
include — (a) isolation and purification of proteins, (b) hydrolysis of proteins to cleave into smaller fragments, and (c) sequencing of amino acids in a protein molecule.

**Extraction of Proteins**

Proteins to be isolated are sourced from microbes, plant or animal cells, or animal tissues. The source material impacts the contaminants present. If the protein is extracellular, one has to segregate the cellular components, whereas for intracellular proteins one has to resort to disruption of cells depending upon their nature. Animal cells are easier to break than plant cells, as animal cells do not have a cell wall. Bacterial cells are very small and need high pressure techniques for disruption. A variety of physico-chemical techniques are used which are described in chapter 33. The ability to detect and quantify the total protein concentration and the target protein levels is an essential requirement for purification and characterization of any protein.

**Protein purification**

Steps in protein purification and scheme of purification are given in Fig. 10.2 and Table 10.1.

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Total protein (mg)</th>
<th>Activity (units)</th>
<th>Specific activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude extract</td>
<td>20,000</td>
<td>400,000</td>
<td>200</td>
</tr>
<tr>
<td>Precipitation by ammonium sulphate</td>
<td>5000</td>
<td>300,000</td>
<td>600</td>
</tr>
<tr>
<td>Precipitation (pH)</td>
<td>4000</td>
<td>1000,000</td>
<td>250</td>
</tr>
<tr>
<td>Ion-exchange chromatography</td>
<td>200</td>
<td>800,000</td>
<td>4000</td>
</tr>
<tr>
<td>Affinity chromatography</td>
<td>50</td>
<td>750,000</td>
<td>15,000</td>
</tr>
<tr>
<td>Size-exclusion chromatography</td>
<td>45</td>
<td>675,000</td>
<td>15,000</td>
</tr>
</tbody>
</table>

For enzyme proteins, their activity is related to catalysis, a given sample is capable of producing. Specific activity is the total activity divided by the amount of protein, and is a measure of a useful purification step. A decrease in protein activity indicates that the purification step is counter-productive. Purification data is usually collected in a tabular form.

**Concentration of Proteins**

Protein concentration is necessary for assessment of purity and activity measurement. Chemical assay methods like UV spectrophotometer (at 280 nm) is a simple and fast, non-destructive process for monitoring concentration of protein in a given sample. Bioassays are sometimes more sensitive than chemical assays. It is necessary to use bioassays when a protein of biological interest like insulin is involved.

Protein recovery of intracellular proteins is normally done by filtration or centrifugation, and subsequently removing other contaminants, like nucleic acids, lipids etc. Aqueous two-phase partition technique is used.

**Aqueous Two-Phase Partition**

Two polymers typically dextran (an α-1, 6-linked glucose polymer with 5% α-1, 3 links), and
polyethylene glycol (PEG) are dissolved in water in appropriate proportions (6% PEG and 8% dextran). Two phases develop, a dextran phase and a PEG-rich phase on top. Proteins distribute themselves between the two phases. One may need to remove or destroy lipids and nucleic acids of the cell homogenate existing as contaminants, which interfere with subsequent purification steps. The lipid layer can be removed if the solution is passed through glass wool or a cloth of fine mesh size. Nucleic acids can be removed by precipitation or by treatment with nucleases. General guidelines to be observed while handling proteins are:

- Maintain pH within the permissible limit
- Avoid conditions, such as heat, that can destabilize proteins
- Use of inhibitors is suggested to prevent action of proteolytic enzymes
- Avoid agitation or chemicals which denature proteins
- Processing time to be minimized

**Hydrolysis of Proteins: Fragmentation**

After obtaining the purified protein, its molecular mass is to be determined. Mass spectrometry is the method which is generally followed. Mass spectrometers consist of three basic parts: an ion source that creates charged molecules in the gas phase; a mass analyzer that uses a physical property, e.g., time-of-flight, to separate ions; and a detector.

Another parameter is the estimation of molecular weight of protein, which can be done by SDS-PAGE or gel filtration. (For details, see chapter 33).

Before commencing primary structure determination of protein, it is to be hydrolyzed or cleaved into smaller fragments. There are two methods available: chemical and enzymatic. Proteins that contain disulphide bridges have to be treated for their removal by 2-mercaptoethanol or dithiothreitol (Cleland reagent). The disulphide bonds are broken and the S is converted to –SH (sulfhydral groups), which need be alkylated by iodoacetic acid to prevent reoxidation.

Cleavage of protein by chemicals is performed by cyanogen bromide, which cleaves peptide chain at the carboxyl side of methionine or tryptophan; by hydroxyl amine cleaving arginine-glycine bonds; and by 2-nitro-5-thiocarbanobenzoate that cleaves cysteine residues. Chemical cleavage alone is not sufficient for the desired purpose, hence cleavage by enzymes is resorted too.

**Amino Acid Composition**

The number of each type of amino acid in a protein molecule can be determined by amino acid analysis. The purified protein sample is hydrolyzed by heating in 6 M HCl at 110°C for 24 h in an evacuated sealed tube. The resulting mixture is called hydrolysate containing amino acids. The hydrolysate is subjected to ion exchange chromatography using a column of sulphonated polystyrene to isolate individual amino acids. The isolated amino acids are then identified and quantified with the help of ninhydrin reaction. Since the hydrolysate contains—amino acids, they produce blue colour, except proline which is an imino acid and produces yellow colour. The amount of each amino acid in an unknown sample can be determined by absorbance (optical density) and comparing with that of known amount of standard sample of amino acid. A highly sensitive detection system has been developed, using fluorescein which reacts with the amino group of amino acid forming a fluorescent product. This method gives us the number of amino acids present in a protein, but not its sequence.
10.4 PROTEIN SEQUENCING STRATEGY

Fredrick Sanger developed a method of protein sequencing, because it is the sequence which determines protein specificity. He showed that proteins are linear polymers of amino acids linked by peptide bonds. Sequencing requires a pure sample of protein. Sanger received Nobel prize twice for his work on proteins (see chapter 33 for Sanger’s reagent).

Edman’s Degradation

Pehr Edman developed a process which is now used in Edman Sequenators, used for protein sequencing. It can determine the first 10 or more residues using only a few picomoles of protein when 50-60 residues are available. However, the necessary condition in this method is that the protein must have a free amino terminal. The N-terminal residue is labelled and then it is cleaved from the rest of the peptide without breaking peptide bonds. In Edman’s degradation, one residue at a time is sequentially removed from the N-terminal end and identified. The amino terminal group is un-charged and therefore it reacts with phenyl isothiocyanate to form a phenylthiocarbamoyl derivative. It is then released from rest of the protein in the form of a cyclic compound, called phenylthiohydantoin amino acid in mild acidic conditions. However, the remaining protein molecule remains intact with a free amino terminal for another cycle of phenyl isothiocyanate reaction. The cyclic phenylthiohydantoin derivative of amino acid can be identified by high pressure liquid chromatography. In this way, each amino acid can be identified one by one, requiring lot of time. Now the sequencing technique has been automated.

The problem arises when the protein molecule consists of higher number of amino acids, say a few hundred or more. In order to sequence a larger protein, the initial step is to break it into smaller fragments to carry on Edman degradation.

Fragmentation of Amino Acids

To obtain smaller fragments of 20–100 residues, the first step is to digest protein molecule with endopeptidases having specific cleavage property. The enzyme trypsin is highly specific for positively charged residues and it cleaves on the C-terminal side of arginine and lysine. Similarly, chymotrypsin cleaves the peptide chain where C-terminal is contributed by aromatic amino acids phenylalanine, tryptophan and tyrosine. Pepsin cleaves peptide bonds where N-terminal is donated by tryptophan, tyrosine, phenylalanine, leucine, aspartic acid and glutamic acid. These peptide fragments obtained by enzymatic cleavage are separated by ion-exchange chromatography and then sequenced by Edman degradation.

Another method of partial hydrolysis of protein molecule is by exopeptidases which are specific for C-terminal cleavage. Carboxypeptidase A cleaves all C-terminal bonds of tyrosine, tryptophan and phenylalanine, whereas Carboxypeptidase B cleaves C-terminal bond of arginine and lysine (Table 10.2). Carboxypeptidase C can cleave all free C-terminal residues.

Although amino acid sequence of each fragment can be known, but the order of fragments in the polypeptide chain cannot be ascertained. Hence, it is necessary to generate overlapping fragments from the original protein. Different chemicals or enzymes are needed for the purpose. Fragments generated by trypsin and chymotrypsin may be used. Chymotryptic peptides will overlap one or more tryptic peptides to establish sequence of fragments (Fig. 10.3).
### Table 10.2 Scheme of enzymatic cleavage of polypeptides

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Site of action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trypsin</td>
<td>Cleaves peptide bonds where C-terminal is contributed by lys and arg</td>
</tr>
<tr>
<td>Chymotrypsin</td>
<td>Cleaves peptide bonds when C-terminal is donated by phe, trp, and tyr</td>
</tr>
<tr>
<td>Pepsin</td>
<td>Cleaves peptide bonds where N-terminal is donated by trp, tyr, phe, leu, asp, and glu</td>
</tr>
<tr>
<td>Carboxypeptidase A</td>
<td>Cleaves C-terminal residues of tyr, trp, and phe</td>
</tr>
<tr>
<td>Carboxypeptidase B</td>
<td>Cleaves C-terminal bond of arg and lys</td>
</tr>
<tr>
<td>Carboxypeptidase C</td>
<td>Cleaves all free C-terminal residues</td>
</tr>
</tbody>
</table>

**Fig. 10.3** The sequence of peptide fragments can be determined by producing overlapping fragments. Solid triangles show the sites of cleavage by trypsin and chymotrypsin

---

### 10.5 MASS SPECTROMETRY IN PROTEIN IDENTIFICATION AND SEQUENCE ANALYSIS

Technological advances in proteomics have led to development of high throughput methods, which are comparatively more efficient and accurate than the experimental methods. Edman degradation method was the method of choice until two decades ago, but mass spectrometric methods for protein identification and sequencing are now being favoured. The mass spectrometer separates proteins according to their mass-to-charge ($m/z$) ratio. The molecule is first ionized and the ion flights are observed in *vacuo* in order to avoid collisional scattering. The process of ionization forces them to move towards the analyzer because of the ionic charge. Ionization can be performed by techniques like MALDI and ESI. MALDI is *matrix-assisted laser desorption ionization* and ESI stands for *electrospray ionization*. The ion is then propelled across the analyzer in an electric field that resolves each ion according to its $m/z$ ratio (Fig. 10.4a and b).

Mass spectrometers consist of an ionization device, a mass analyzer that separates the mixture of ions, and a detector. The material to be analyzed is transferred into the gas phase and subjected to ionization, making use of MALDI and ESI methods.
Ionization in MALDI is done by having a purified protein suspended or dissolved in a crystalline matrix of small, organic, UV absorbing molecules. The crystal absorbs energy at the same wavelength of the laser that is used to ionize protein. The laser energy strikes the matrix to cause rapid excitation of the matrix and subsequent passage of matrix and analyte ions into the gas phase. The principle ion detected using threshold laser intensity for MALDI is \([M+H]^+\), although signals for multiple charged ions and
oligomeric forms of the analyte may be seen, especially for large proteins. The ionized protein is accelerated by an electrostatic field and expelled into a flight tube. As it exits the flight tube, the mass analyzer is encountered. The analyzer is often a time-of-flight (TOF) analyzer, which is based on the principle that when accelerated by application of constant voltage, the velocity with which an ion reaches the detector is determined by its mass. MALDI is able to analyze proteins down to femtomole quantities. Also MALDI is able to tolerate small amounts of contaminants; therefore sample preparation is not as tedious as with ESI mass spectrometry. The information obtained from MALDI analysis is subjected to a database search.

The sample to be analyzed is spotted onto a plate, dried and then put inside the machine. With MALDI-TOF the sample needs to be ionized and vaporized and the laser supplies the energy. The sample is placed in a matrix (a chemical) and placed on a metal plate. After firing the laser, the matrix absorbs the energy and the sample is ionized rapidly. It is then accelerated through an electric field and allowed to pass through the analyzer in vacuum. The time-of-flight is measured and mass/charge ratio is calculated for the peptide fragments (Fig. 10.5).

**Peptide Mass Fingerprinting (PMF)**

This is another method of protein identification. In this method, 2-D gel electrophoresis is used for protein separation. The separated spots are obtained from the gel and then identified by PMF. The technique is based on the use of a proteolytic enzyme to digest the protein into smaller peptides. The most commonly used enzyme is trypsin, which cleaves lysine and arginine sites. When the digestion is complete, a set of peptides are produced of varying masses that are unique to that protein. The mass of each peptide will be the sum of amino acids present, inducing any modifications that amino acids might have undergone. Once the set of peptides have been obtained, one has to search for peptide sequences.
REFERENCES AND RECOMMENDED READINGS


PROBLEMS

1. Explain different types of covalent and noncovalent bonds that exist in proteins.
2. A polypeptide has the following amino acid sequence. When it is treated with chymotrypsin, how many peptide fragments will be formed? What will be their amino acid sequence?
   Val–Phe– Asp– Lys– Gly–Phe– Val–Glu–Arg
3. How can you determine the primary structure of a protein?
4. Describe the steps in purification of proteins.
5. What should be the strategy to sequence amino acids in a polypeptide? Why Edman degradation is preferred over Sanger’s method?
6. Explain peptide mass fingerprinting.
7. Give brief account of the following:
   (a) Mass spectrometry and its purpose
   (b) MALDI–TOF