Commercial Importance

Some estimates suggest that the human body may contain over two million proteins. These are coded by only 20,000 - 25,000 genes. The total number of proteins found in terrain biological organisms is likely to be more than ten million. This class of molecules is very complex. Each cell in the human body contains thousands of different proteins. Thus, proteins are the most abundant and diverse class of biomolecules. Only a small percent of the proteins have been analyzed for their structure and fully sequenced. The detailed secondary and tertiary structures have been established for only a few thousand proteins. As of April 2010, Protein Data Bank (PDB) holdings list 52464 X-Ray structures of proteins. The known chemistry clearly points to several well-defined structural patterns in proteins. Proteins are made up by linear condensation of different amino acids. The carboxylic acid group of one amino acid forms an amide linkage with the ω-amino group on the second amino acid. The process continues up to the desired chain length. The number of amino acids in a protein could range from two to several thousand. The largest known proteins are the titins, a component of the muscle sarcomere, with a molecular mass of 3,816,188.13 Da, and a total length of 35,213 amino acids.

When the length of such a chain is up to 100 amino acids, they are called Media:peptides. Up to a length of 20 or so, they are referred to as small peptides. The molecules having more than 100 amino acids are called proteins. These numbers and names like peptides, polypeptides etc., are very arbitrary. Proteins play several vital roles in the biochemistry called life. Proteins were first described by the Dutch chemist Gerhardus Johannes Mulder and named by the Swedish chemist Jöns Jakob Berzelius in 1838. The central role of proteins in living organisms was however not fully appreciated until 1926, when James B. Sumner showed that the enzyme urease was a protein.

Peptide bond

The amide bond that connects two consecutive amino acid segments in peptides and proteins is called the Media:Peptide Bond. This is just a special name given to these amide bonds. Once such a chain is constructed, the chain would have two different terminals viz., the N-terminal bearing the free (or protected) amine group and the C-terminal bearing the carboxylic acid (or a suitable derivative like ester or even an amide, other than a peptide bond) as end group. The convention used amongst chemists is to place the N-terminal on the left-hand side of the picture and place the C-terminal at the right-hand end of the structure, irrespective of the length of the chain. Long chains are folded either upwards or downwards. This allows the stereochemistry at the asymmetric centers to be projected in the same way, avoiding confusions. Using this arrangements, the amino acids are generally written with either the three-letter abbreviation or
the one-letter abbreviation (see Table 1.5 for the abbreviation used for the proteinogenic amino acids). For new amino acids, the chemists use new abbreviations to suite the nomenclature. A few examples are provided below to illustrate these points. You could easily appreciate the convenience of using such abbreviation while writing longer chains. The approved nomenclature for peptides / proteins is simple and straightforward. The ending ‘-ine’ for all amino acids (except the last amino acid at the C-terminal) is replaced by ‘-yl’ and these names are written in the same sequence as the assigned structure, without any spacing. The full name of the last amino acid is written in the last. The entire name is written as one word. The derivatives at the terminals are prefixed and surfixed at the N- and C- terminal respectively. A few examples are given below (Fig 3.1).

**Fig 3.1: Nomenclature in simple peptide sequences.**

A vertical bond with appropriate abbreviations on the type of modifications would show the derivatives on the polar side chains (Fig 3.2).

**Fig 3.2: Indicating side-chain protections – the vertical bond on Glu- and OMe means the carboxylic acis side chain of Glutamic acid has a methyl ester.**
Ramachandran Plot and its significance:

The diagram below shows one amino acid unit in a peptide chain (Fig 3.3). Note that the carbonyl at the C-terminal and the carbonyl at the N-terminal are ‘trans’ i.e. the C=O groups are directed in opposite directions to the direction of the peptide chain, which is oriented north-south in this diagram. This trans- orientation is preferred in most of the units to minimize the dipole repulsions, though a small percent of the peptide units could be ‘cis’ due to steric constraints. The Cα – CO bond of the amino acid unit is called the ψ-bond (psi), while the Cα – N bond is called the φ-bond (phi). The bond between the C-terminal carbonyl and the nitrogen of the next amino acid unit is called ω-bond (omega) (Fig 3.4).
Linus Pauling and Robert Corey analyzed the geometry and dimensions of the peptide bonds in the crystal structures of proteins. Their results are summarised in this diagram (Fig 3.5). The consensus bond lengths are shown in Angstrom units. And the bond angles are given in degrees.

Note that the orientation of the Ca – CO bond and the next N - Ca bond could be either cis- or trans-. These peptide bonds are mostly in the trans configuration since it is more favourable than cis-. The amide bond exhibits hindered rotation across the N – CO bond. The peptide bond (like any amide bond) exists as a resonance hybrid of two canonical structures (Fig 3.6). Due to this resonance effect, the lone pair on nitrogen partially overlaps with the carbonyl carbon. This introduces a partial double bond character across the N – CO bond single bond, leading to the observed hindered rotation. Depending upon the energy of such hindered rotation, one could expect
cis- / trans- isomerisation across the amide bond (Fig 3.7). In complex systems like the proteins, where several other steric parameters exist, the molecule could freeze in one of the conformations. The cis- configuration is sometimes found to occur with proline residues.
G.N. Ramachandran et al., looked into the hindered rotations around the φ-bond and the ψ-bond, considering the van der Waal radii of the concerned atoms as hard spears (i.e. region that are not allowed). The resulting φ / ψ – Ramachandran G.N. et al., J. Mol. Biol., 7, 95 (1963); Adv. Protein Chem., 23, 283 (1968).

plot is now known as Media:Ramachandran Plot (Fig 3.8). The figure below shows some important features and information that could be drawn from a Ramachandran map. Let us now look at some of these complex sub-structures that are found in peptides and proteins.
Structure of Protein: Primary, Secondary and Tertiary Structures

Primary Structure:

The fact that proteins are made up of a long linear chains of amino acids is only a part of the story. This information tells us the sequence in which the amino acids are condensed to make the chain. This information is called the Primary Structure of a protein molecule. The hindered rotations and hydrogen bondings, mainly between the oxygen of the amide carbonyls and a distant CONH would constrain the linear molecule to bend into well defined Secondary Structures.
Fig 3.9: Primary Structure indicates the sequence of amino acids in the peptide (protein) chain.

Secondary Structures:

As discussed above, the ϕ-bonds, ψ-bond and ω-bonds constrain these bonds to bend away from linearity.

**Media:** α-Helix: In 1951, L. Pauling proposed α-Helix (Fig 3.10) as a stable structural unit in proteins. Pauling’s model suggested repetitive hydrogen bondings between NH and CO, four residues apart. Later such α-Helix units were observed in proteins. In principle, the helixes could be right handed or left handed. For L-amino acids, left-handed helixes are energetically disfavoured and are seldom observed. Helical structures could occur anywhere in the protein chain. We describe the H-bonding as occurring between consecutive ith and i + 4th residues (5→1), resulting in 13 atoms being held by one H-bonding. Other modifications of this structure are also known. A study by Barlow et. al., of 54 proteins (whose X-ray atomic coordinates are known) indicate that 35% of the residues are helices, out of which 80% are α-Helices.

β-Pleated Sheets (β-strand): These are the second major structural elements found in proteins. In these structures, the consecutive amino acid units do not ‘turn’ as in helixes. They remain zig-zag and grow in a linear fashion. Two such linear chains, connected by a loop of several amino acid units, could orient parallel to each other. The NH and C=O moieties that face each other are held by hydrogen bonding. Thus, a series of hydrogen bondings hold the strands in a parallel fashion. This arrangement of the strands could be of two types. The chains could grow in the same direction or in opposite directions. When the strands grow in the same direction they are called Parallel Sheets. When the strands grow in opposing directions, they are called Anti-parallel Sheets (Fig 3.11). Such a β-strand is a stretch of amino acids, typically 5–10 amino acids long, whose peptide backbones are almost fully extended.
Note that the two types of β-sheets show marked differences in H-bonding patterns. β-sheet is a repeating secondary structure in proteins, compromising 20-28% of all residues in globular proteins. The beta sheet is often called pleated because sequentially neighboring carbon atoms are alternately above and below the plane of the sheet, resulting in a pleated appearance (Fig 3.12).
**Reverse turns:** When only two consecutive peptide units bend in the same direction (as in a Helix), while the peptide units before and after these bonds maintain the linear (actually zigzag) structure, the protein is said to make a **U-Turn** at these two peptide units. The peptide chains on either side of the bend run antiparallel to each other. They are further classified on the basis of $\phi$, $\psi$, and $\omega$ bond angles. Two types of reverse turns are shown below (Fig 3.13).

---

**Introduction to protein structure, 2nd Ed, By C. Branden and J. Tooze, Garland Publications**

---

**Fig 3.12: The pleated arrangements in parallel**

---

**Fig 3.13: Two different types of U-Turns in peptide structures.**

**Super-secondary Structures**

In proteins, these secondary structures could occur in different combinations to make **Super-secondary Structures**. Some of the commonly occurring **folding motifs** are shown below (Fig 3.14). Note that these drawings depict the protein chain as a ribbon. They are called ribbon drawings. Such drawings are least clustered (with chemistry details) and focus totally on the structural patterns. These are sometimes referred to as cartoonic drawings.
Tertiary Structures:

When several super-secondary structures are incorporated along the length of the entire peptide molecule, the molecule is no longer a random coil structure (as seen in most polymers). As shown below, the actual shape is a coil, in which the segments are organized in well-defined secondary and super-secondary structures throughout the length. This is called Tertiary structure of proteins. While representing ‘tertiary structure of a protein, it is customary to use ribbon drawings. Some such well-structured tertiary structures of proteins are shown below (Fig 3.15).
Other bonds in tertiary structures:

In addition the H-bonds, there are other types of bonds that hold super-secondary and tertiary structures in proteins together.

**Disulphide Bridge:** Two –SH terminal on the side chain of cysteine residues could be oxidized to a disulphide linkage. Unlike the H-bondings that are weak, these disulphide bridges are strong covalent bonds that are cleaved only by reductive enzymes or reagents.

**Metal chelates:** Metal ions like calcium and zinc could form different types of chelated structures with amide bonds and other functional groups. These sites are often very critical elements in biochemical reactions.

**Salt Bridges:** Oppositely charged side chains e.g., COO- / - +NH3 could hold two distant segment together. These are important linkages in tertiary and quaternary structures of proteins.
Hydrophobic Interactions: These are weak bonds. When several such weak bonds occur, they become strong enough to hold large protein segments together. The non-polar groups mutually repel water and other polar groups and results in a net attraction of the non-polar groups towards each other. Alkyl groups on Ala, Val, Leu etc., interact in this way. In addition, benzene (aromatic) rings on Phe and Tyr can “stack” together. In many cases, this results in the non-polar side chains of amino acids being on the inside of a globular protein, while the outside of the proteins contains mainly polar groups. These are important regions on the surface of proteins, where contact occurs between different biomolecules.

Quaternary Structures:

Several proteins have more than one primary chain-clusters in their structures. They are often inactive as monomers. Indeed each monomer has well defined secondary and tertiary structures. However, the bioactive protein is complete only when the different components are assembled together. Such proteins are called **oligomers** and the individual monomers are called **subunits**. They could be dimmers, trimers, tetramers or hexamers. **Hemoglobin**, the protein that carries oxygen in our blood, is a tetramer. This structure is complete only when the iron atom of **Heme** (also called iron photoporphyrin IX) is ligated with the histidine of the protein. This iron is further complexed with oxygen molecule (or CO2). The structure of hemoglobin and the Heme unit are shown below (Fig 3.16).
Insulin molecule

The primary structures of Chain A and chain B of insulin are shown below (Fig 3.17). Note that the two chains are linked by two disulphide linkages (Inter-chain disulphide bridges). In addition, the Chain A has one intra-chain disulphide bridge as well. Insulin molecules have a tendency to form dimers in solution due to hydrogen-bonding between the C-termini of B chains.

Fig 3.17: Primary structure of insulin molecule

The bioactive ‘insulin’ is actually a hexamer. In the presence of zinc ions, insulin dimers associate into hexamers. The hexamer with the central Zinc ion is shown below (Fig 3.18).
Sequence determination

The primary structure of a peptide or protein is first determined. The secondary, tertiary and quaternary structures are later determined using NMR and X-ray diffraction techniques. In this course let us have a brief look at the determination of the primary structure of a protein. Determining the sequence in which the amino acids are arranged in a peptide or protein is known as Determination of Protein or Peptide Sequence.

What are the amino acids present?

The purified protein is first digested with acid to hydrolyze all the peptide bonds. This procedure causes all the peptide / amide bonds to cleave to give a mixture of amino acids of which the molecule is made. Using an amino acid analyzer, the type of amino acids and their relative abundance could be accurately determined using chromatography procedures. The hydrolysis procedure destroys most of the tryptophan. It also partly destroys cysteine and cystine. Serine and threonine are slowly destroyed. The side-chain amide bonds are also cleaved. In spite of such limitations, this procedure is often the first step in structure elucidation.

Mass Spectroscopy techniques are now refined to a great level. MS data provide substantial inputs into the molecular weight, the amino acids present and mass of several fragments. All these inputs together help
us to solve the structure of proteins. The main advantage of this technique is that only a few micro-
grams are needed for MS analysis.

Alkaline hydrolysis destroys some of the sensitive amino acids but does not affect tryptophan. However, this procedure causes considerable racemisation. Notwithstanding such limitations, these procedures give the type of amino acids present and their relative abundance.

The process from here on is very complex.

The Problem of Sequencing:

Let us consider a small tripeptide. From hydrolysis we could find out that the peptide G-A-V has glycine, alanine and valine in the ratio 1:1:1. However, these three amino acids could be assembled in (3!) ways. Thus six structures are feasible.

GAV, GVA, AGV, AVG, VGA, VAG

As the number of amino acids in the sequence increases, the situation becomes very complex. For example, for a peptide made up of four amino acids A, B, C and D, there are (4!) i.e. 24 possible combination of structures. How do we go from here on?

1. If we have a reagent that is specific to one of the terminal group we would know the end terminals.
2. Starting from one end, if would cleave the amino acids one by one and identify them, the sequencing becomes a simple job.
3. For longer sequences as in proteins, the given protein could be first hydrolysed to smaller fragments. For each fragment, we could find out the amino acids present at the C-terminal and N-terminal by independent procedures. The sequence of each of these small peptides could then be determined. With these inputs, we could logically assemble the fragments using a deductive procedure called ‘overlap method’.

Cleavage of disulphide linkages:

The first step is to reduce the disulphide linkages, if any. A common procedure is to reduce with Media:mercaptoethanol and then methylate the S – H to S – Me. This protection step avoids further oxidation to disulphides during the degradation procedures (Fig 3.19).
Fig 3.19: Cleavage of a disulphide bond and protection of the liberated –SH unit

Cleaving the terminal groups one by one:

N-Terminal analysis:

The most widely used procedure is the Media:Edmund Procedure at C-terminal. The procedure relies on the ability of phenylisothiocyanate (PITC) to react selectively with free amino groups to form thioureas. In the given procedure, the thiourea derivative is treated with HF to give a phenylthiazolinone derivative (PTH). A peptide with one less amino acid unit bearing a free –NH2 terminal is released. The procedure is depicted below for a tetrapeptide (Fig 3.20). The two-step procedure yields the N-terminal amino acid derivatives one after another. The thiazolinone derivatives of all known amino acids are matched with the degradation product obtained. Thus the sequence is delineated. The procedure is slow and needs large quantities of peptides. Due to accumulation of byproducts, the procedure has limitations. In an automated sequencer, the method has been applied successfully for peptides having upto 50 amino acid units.
A procedure called Media:DNP Method (1945) identifies the N-terminal in a given peptide. The peptide with a free N-terminal is treated with 1-Fluoro-1,4,-dinitrobenzene (FDNB) under mildly basic conditions. Under this condition, the asymmetric centers and peptide bonds are unaffected. The resulting N-DNP derivative is hydrolyzed with dil. acid. The mixture is separated by chromatographic procedures and the N-DNP derivative is identified from known physical parameters (Fig 3.21). Unlike the Edmund procedure, this procedure could be applied only once on a peptide.
In subsequent modifications dansyl chloride was introduced as a reagent for the N-terminal. The structure of this reagent is given below (Fig 3.22). This procedure is now called the Media:Dansyl Method. This modification gives derivatives that are Fluorescence active and are therefore amenable to micro-scale determinations. The limitation of both these methods lies in the fact that the reagents attack the side-chain amino groups as well and are therefore not selective to the peptide chain.
Determination of the C-terminal:

In 1956 Akabori et al., reported a procedure that targeted the C-terminal of the unprotected peptides. They treated a peptide with anhydrous hydrazine at 100°C.


All the peptide bonds were attacked and the amino acids converted to the hydrazide derivative, except the C-terminal acid. The mixture was eluted on cation-resin column. The free amino acid was eluted selectively and identified (Fig 3.23).
Fig 3.23: Media:Akabori’s hydrazide procedure for C-terminal.

Like the DNP method on N-terminal, this procedure could be used only once on a peptide.

Another procedure is to reduce the peptide with lithium borohydride reagent that is specific to carboxylic acids. The COOH group is selectively reduced to CH2OH. On hydrolysis of the peptide bonds, the aminol at the terminal of the chain is separated by extraction and identified. Here again, this reagent attacks the side chain acid groups as well. However, hydrolysis of the peptide chain leads to amino acids at these spots as well.

The C-terminal analysis could also be carried out with Media:exopeptidase enzymes. Exopeptidases are selective to peptide groups at the C-terminal. Media:Carboxypeptidase A cleaves the C-terminal amino acid unit only. It does not work if the terminal unit is arginine or lysine. On the other hand Media:Carboxypeptidase B works selectively on arginine and lysine.

Partial Hydrolysis method:

Once the N- and C- terminal units are identified for a protein, the molecule is subjected to partial hydrolysis procedures. Here, the hydrolysis reactions using acids are stopped at various time intervals. This gives a mixture of smaller peptides. These are separated by chromatography techniques. Since small peptides are easy to analyze by Edmund Method, the sequences of these fragments are easily established. Hydrolysis reactions are random processes and therefore do not give the same fragments when experiments are repeated. This inconsistency is an advantage for sequence determination studies. A large number of different fragments are obtained by such repetitions. Using ‘overlapping procedure’ the sequence could be determined. Let us consider a simple example of a peptide, A-B-D-D-C-A-E . Total
Hydrolysis tells us that the peptide is made up of amino acids 2A, B, C, 2D, and E. Partial hydrolysis studies give the following smaller peptides.

The N-terminal is known as ‘A’ . The fragments are overlapped as follows:

This procedure works well for small peptides. Using regio-specific media: Endopeptidases, larger proteins are fragmented at specific points. These peptidases are listed in Table 3.24.

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>C-terminal Site</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trypsin</td>
<td>C-end of Arg and Lys</td>
</tr>
<tr>
<td>Chemotrypsin</td>
<td>C-end of Typ, Phe, Tyr</td>
</tr>
<tr>
<td>Elastase</td>
<td>C-end of Gly, Ala</td>
</tr>
</tbody>
</table>

The enzyme catalyzed fragmentation procedures provide fewer fragments, whose C-terminals are known. This simplifies the procedure.

Another specific reagent is cyanogens bromide. This chemical is specific to a methionine residue. While the enzymes are sensitive to the shape of the site and thus do not work near the specified sites (for example, if proline is present), cyanogens bromide has no such limitations. The mechanism of cleavage is as follows (Fig 3.25).
Thus, the procedure fragments the protein at methionine residues only.

The last step would be to identify the location of the disulphide linkage. This is done by partial hydrolysis of the whole peptide. The sequence on the fragments bearing the disulphide links are identified as discussed above,

The presence of dimers, trimers etc., in protein structures present a complex problem in sequence determination. Such molecules give multiple ‘end groups’ pointing towards such noncovalently bonded complex structures. Advances in Mass Spectroscopy have provided a convenient tool, not only to identify primary structures (by study of the fragmentation patterns as we saw in partial hydrolysis and mass fragments of individual amino acids), but also by looking at dimeric, trimeric structures in mass spectra. Techniques like. Electron Spray Ionisation (ESI) and MALDI are
convenient tools for such delicate biomolecules.