3.20. STRUCTURE OF PEPTIDES AND PROTEINS

The structures of proteins are quite complex. Their complete structures are usually discussed at four different levels, *i.e.*, primary, secondary, tertiary and quaternary structures, each level being more complex than the previous one.

(i) **Primary structure.** Proteins may have one or more polypeptide chains. The primary structure of a protein refers to the covalent structure including disulphide bridges of each polypeptide chain. *It simply refers to the sequence in which the various amino acids present in a protein are linked to one another* (Fig. 3.4).

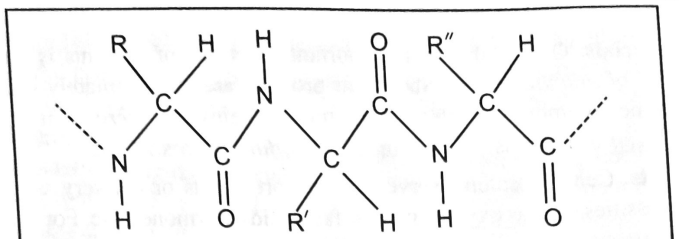


FIGURE 3.4. Primary structure of a protein.

The primary structure of a protein or a polypeptide can be determined by finding answers to the following three questions :

(*i*) *What different amino acids are present* ?

(*ii*) *How many moles of each amino acid are present* ?

(*iii*) *What is the sequence in which the different amino acids occur in the polypeptide chain ?*

The determination of primary structure of any protein involves the following two steps.

1. Determination of amino acid composition. To answer the first two questions, the protein is completely hydrolysed* to its constituent amino acids either by heating with 6N HCl at 373–393 K for 20-24 hours or by enzymes. Alkaline hydrolysis is generally not used since it causes racemization. The resulting mixture of amino acids is separated either by ion-exchange chromatography or by gas chromatography of their methyl esters.

In ion exchange chromatography, the mixture of amino acids is dissolved in an aqueous buffer and allowed to pass down the column. As each amino acid leaves the chromatographic column, it is mixed with **ninhydrin** which produces purple colour with amino acids. The purple colour is monitored by a spectrophotometer. The intensity of the colour is proportional to the amount of the amino acid present. It is plotted against retention time (*i.e.,* time of elution) when a **chromatogram** (Fig. 3.5) consisting of a series of peaks of different sizes is obtained. The positions of peaks are characteristic of individual amino acids and the areas under the peaks correspond to their relative amounts.

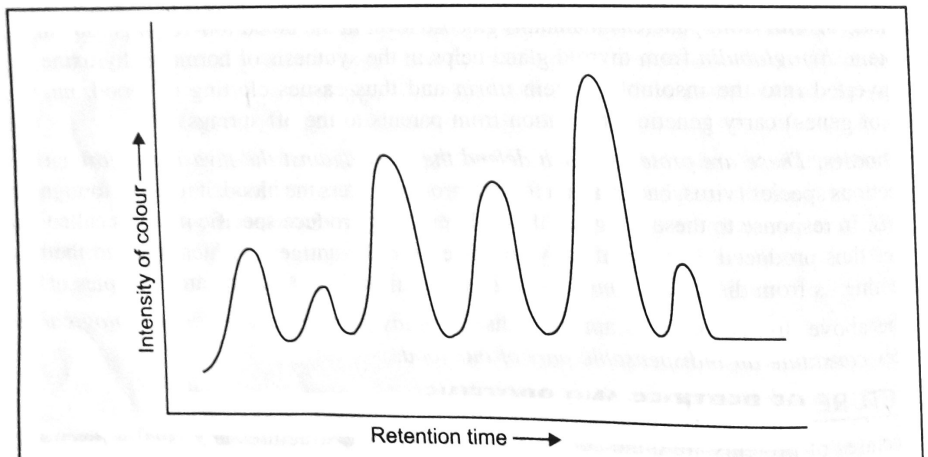


FIGURE 3.5. Chromatogram showing different amino acids and their relative amounts.

*If the protein contains disulphide bridges, the protein is first treated with performic acid and then subjected to hydrolysis. During this oxidation, cysteine bridges are oxidised to sulphonic acid units.

From the weight of each amino acid, the number of moles of each amino acid is calculated. This gives us the relative numbers of different amino acids present in the given protein. The molecular weight of the protein in them determined either by a chemical method or by one of the many physical techniques such as osmotic pressure, light scattering measurements, ultracentrifugation, X-ray diffraction, etc. From the molecular weight, the molecular formula and number of moles of each amino acid residue actually present in the protein molecule is then calculated.

The whole process is now automated, *i.e.*, carried out automatically by a commercial instrument called the **amino acid analyser.**

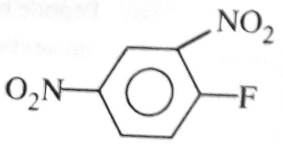
The technique is so sensitive that only 10^{-5} to 10^{-7} g of the peptide is required to carry out amino acid analysis.

2. Sequencing of amino acids. After the identity and amount of each amino acid constituting the polypeptide chain has been found out, the next step is to determine the sequence in which the amino acids occur along a polypeptide chain. This is done by a combination of terminal residue analysis and partial hydrolysis as discussed below.

Terminal residue analysis. Each peptide chain has two terminal amino acid residues, one of these has a free amino group and is called the **N-terminal residue** while the other has a free carboxyl group and is called the **C-terminal residue.** These terminal amino acids are identified by using certain specific reagents. *The whole process consists of treating the given polypeptide with a suitable reagent which reacts either with the N-terminal or the C-terminal amino acid residue to form the tagged peptide. The tagged peptide is then subjected to partial hydrolysis which releases the tagged amino acid which is then identified.* The resulting peptide (shortened by one amino acid) is tagged and subjected to partial hydrolysis again to yield the second *tagged amino* acid which is again identified. This process is repeated number of times till the entire peptide chain is degraded and all the amino acids identified. Depending upon the nature of the terminus, two types of reagents have been developed.

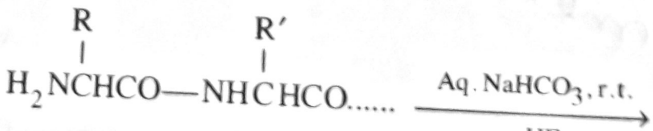
(a) N-Terminal residue analysis. This can be done by the following two methods.

1. Sanger's method. It consists of treating the peptide with *2, 4-dinitrofluorobenzene (DNFB)* also called the **Sanger's reagent** in presence of mildly basic solution (sodium bicarbonate solution) at room temperature. The presence of electron-withdrawing nitro groups at *o*- and *p*-positions activates the fluorine atom towards nucleophilic attack by the $-NH_2$ group of the N-terminal amino acid residue. The resulting 2, 4-dinitrophenyl (DNP peptide is hydrolysed with dil. HCl to the constituent amino acids and the yellow 2, 4-dinitrophenylamino acid (DNP–AA) which is separated and identified.

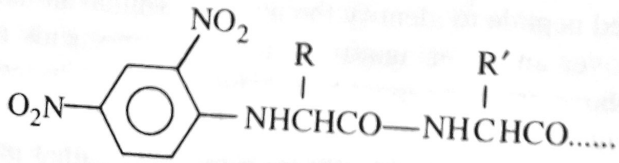


2, 4-Dinitrofluorobenzene
(DNFB)

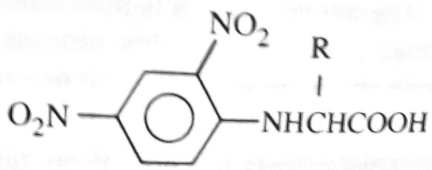
+



Peptide



Tagged peptide



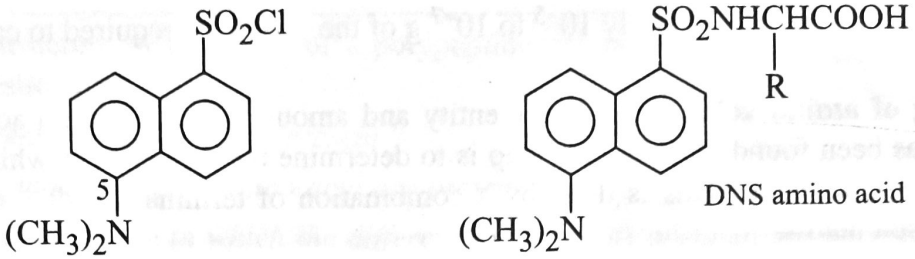
N-(2, 4-Dinitrophenyl)–
amino acid (DNP–AA)

+ Mix. of amino acids
(Untagged amino acids)

DNFB method

Demerits. This method as such is not very useful since only the terminal amino acid is identified. However, if the polypeptide is first hydrolysed into a number of smaller peptides and then each one of these peptides subjected to DNP analysis, it is possible to identify the sequences of all the amino acids in the polypeptide chain.

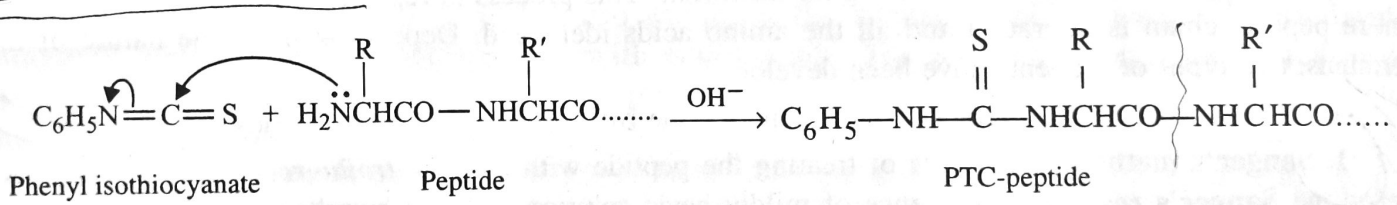
2. Dansyl method. Dansyl method is a recent modification of Sanger's DNP-method. In this method, instead of Sanger's reagent (DNFB), 5-dimethylaminonaphthalene-1-sulphonyl chloride called dansyl chloride (DNS-Cl) is used and DNS amino acid thus formed identified.



DNS amino acid

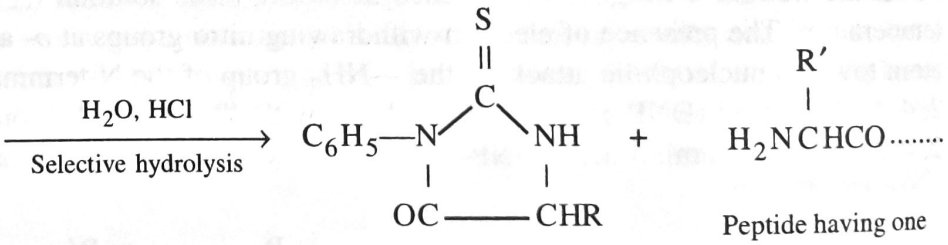
The dansyl method is now being widely used because the dansyl group is highly fluorescent. Therefore, it permits detection and estimation of dansyl amino acids in minute quantities by fluorimetric methods.

3. Edman's method. A more useful and more widely used method for N-terminal analysis is the **Edman's degradation.** It involves the reaction of a peptide with *phenylisothiocyanate* in presence of dilute alkali when the $-NH_2$ group of the terminal amino acid of the peptide reacts to form a N-phenylthiocarbamyl **(PTC)**-peptide. This upon mild hydrolysis with HCl, selectively removes the N-terminal amino acid as the *phenylthiohydantoin* **(PTH)** but the rest of the peptide chain remains intact.



Phenyl isothiocyanate Peptide PTC-peptide

Edman's method



A phenylthiohydantoin (PTH)

Peptide having one amino acid residue less than the parent peptide

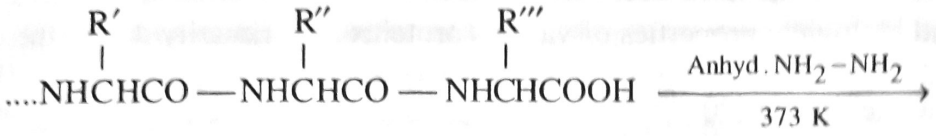
The phenylthiohydantoin thus produced is separated from the degraded peptide and identified chromatographically by comparing it with phenylthiohydantoins obtained from known amino acids. This leads to identification of N-terminal amino acid.

The above process is then repeated on the degraded peptide to identify the new N-terminal amino acid residue. Theoretically, this process can be repeated over and over again till all the amino acids in the peptide chain are identified. In practice, however, the above process cannot be extended beyond 20 residues due to interference of the amino acids produced by slow hydrolysis of the peptide during acid treatment.

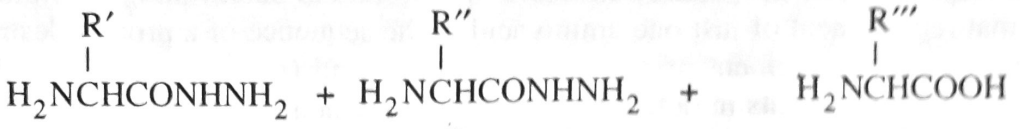
This process has now been automated, *i.e.,* is carried out automatically by a machine called **protein sequenator.** By this machine, it was possible to sequence 40 of 153 amino acid residues present in *myoglobin,* a protein which transports oxygen from lungs to tissues in sea mammals such as whale, seal, walrus, etc.

(*b*) **C-Terminal residue analysis.** The two commonly used methods for C-terminal analysis are described below :

1. Hydrazinolysis method. This is the most widely used method for C-terminal residue analysis in a protein or a polypeptide. In this method, the polypeptide is heated with anhydrous hydrazine at 373 K when all amino acid residues except the C-terminal one is converted into amino acid hydrazides. The mixture of products thus obtained is subjected to chromatography over a column of a strong cation exchange resin. On elution, the strongly basic hydrazides are retained but the free amino acid is eluted. By identifying the free amino acid, the C-terminal amino acid residue of a protein or a polypeptide can be determined.



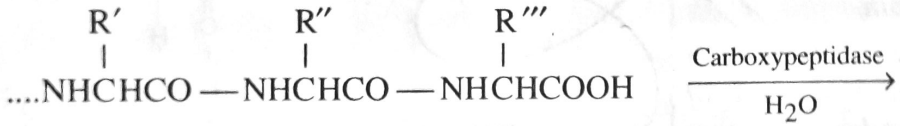
Polypeptide



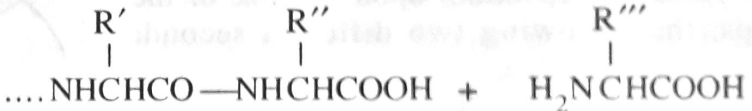
Hydrazides Free C-terminal amino acid

2. Enzymatic method : Selective hydrolysis. Chemical methods for C-terminal amino acid residue analysis have not been very successful. However, it can be carried out enzymatically.

The enzyme *carboxypeptidase* (obtained from the pancreas) is specific for cleaving the peptide linkage next to the free α-carboxyl group in the peptide. The process involves the treatment of a peptide with the enzyme when a free amino acid along with the degraded peptide is formed.



Polypeptide



Degraded peptide C-Terminal amino acid

The amino acid is identified and the degraded peptide with a new C-terminal residue is again treated with the enzyme to yield a second free amino acid and a further shortened peptide. The process is repeated over and over again till the sequence of all the amino acid residues in the protein is worked out.

In actual practice, however, it is not possible to determine the sequence of all the amino acids in a polypeptide chain by stepwise removal of one terminal amino acid residue at a time. Instead, the given polypeptide chain is subjected to partial hydrolysis with an acid or an enzyme and the smaller peptides so formed, *i.e.*, dipeptides, tripeptides, tetrapeptides...etc. are identified by the terminal residue analysis. When enough of these smaller peptides have been identified, it is possible to find out the full sequence of amino acid residues in the entire polypeptide chain by matching a number of such overlapping peptides.

As an illustration of the sequencing procedure, let us consider an extremely simple example of a widely distributed tripeptide, *glutathione* which can have six possible structures.

In order to find out the exact structure, the following procedure is applied.

(*i*) Amino acid analysis shows the presence of three amino acids, *i.e.*, glutamic acid, cysteine and glycine.

(*ii*) The N-terminal analysis by Edman's method reveals that glutamic acid is the N-terminal residue.

(*iii*) Partial hydrolysis gives the following two dipeptides whose structures have been determined by amino acid analysis and N-terminal residue analysis.

Glu–Cys
Dipeptide-I

Cys–Gly
Dipeptide-II

(*iv*) Matching the overlapping dipeptides, the following structure for glutathione is established.

Glu–Cys–Gly

Glutamylcysteinylglycine

(*Glutathione*)

The first ever primary structure of a protein, *i.e.*, insulin* was determined by the British chemist, *Frederic Sanger* and for this work he was awarded the Nobel Prize in 1958. Since then, the primary structures of hundreds of different proteins have been determined.

The different chemical and biological properties of various proteins are primarily due to the differenes in their primary structures. A protein containing 100 amino acids is a very small protein, yet 20 different amino acids can be combined at one time in $(20)^{100}$ different ways to give an equal number of proteins each having its own characteristic properties.

The importance of the primary structure of a protein in determining its biological activity is shown by the fact that replacement of just one amino acid in the sequence of a protein destroys its biological activity. For example, the replacement of one specific amino acid (glutamic acid by valine) in *haemoglobin*, the protein of the blood, results in defective haemoglobin which causes a disease called *sickle cell anaemia*. In the patients suffering from this disease, the defective haemoglobin precipitates causing the cells to sickle and sometimes even makes them burst leading ultimately to death.

–Val–His–Leu–Thr–Pro–[*Glu*]*–Glu–Lys–*

Normal haemoglobin

–Val–His–Leu–Thr–Pro–[*Val*]*–Glu–Lys–*

Sickle cell haemoglobin