Lecture 18: Protein Sequencing

Frederic Sanger first time achieved complete sequence of protein (bovine insulin) in 1953. For his work, he was awarded the Nobel Prize of Chemistry in (1958).

Protein sequencing refers to the techniques employed to determine the amino acid sequence of a protein. There are several applications of protein sequencing, which are:-

a) Identification of the protein family to which a particular protein belongs and finding the evolutionary history of that protein. Function prediction.
b) Prediction of the cellular localization of the protein based on its target sequence (sequence of amino acids at the N terminal end of the protein which determines the location of the protein inside the cell).
c) Prediction of the sequence of the gene encoding the particular protein.
d) Discovering the structure and function of a protein through various computational methods and experimental methods.

Till date several methods have been utilized for protein sequencing. Two main methods include Edman degradation and Mass Spectrometry. Protein sequence can also be generated from the DNA/mRNA sequence that codes for the protein, which has been explained in details in the recombinant DNA section. Here, we have discussed the most important methods used for protein sequencing and the pros and cons of each method.

Edman degradation

Before sequncing procss is initiated, it is necessary to break all non-covalent interation by denaturants (like high concentration of urea or GuHCl). This process will also separate subunits, incase of oligomeric proteins. Ocassionally, subunits of a oligomeric protein are connected by covalent interations. In that case special treatments are required to separate subunits. The protein is treated with Edman's reagent (phenyl isothiocyanate) which reacts with the N-terminal amino acid and under mild acidic condition forms a cyclic compound Phenyl thiohydantoin derivative (PTH–amino acid) of N-terminal amino acid is released. Amino acid of PTH –amino acid derivative is identified by chromatographic

property of the PTH –amino acid derivative. In this process N-terminal aminos acid is identified after first cycle. ***Since this method proceeds from the N terminal residue, the reaction will not work if that N-terminal of a protein is blocked (generally due to post-translational modification)***. After first cycle of the reaction, amino group of the second amino acid is free for reaction with Edman's reagent and at the end of reaction PTH derivative of second amino acid from N-terminal is released. The process continues till end of sequence or a disulfide bond is encountered in the sequence. PTH-cysteine derivative will remain attached with polypeptide and PTH-cystein will not be release (Fig. 1)

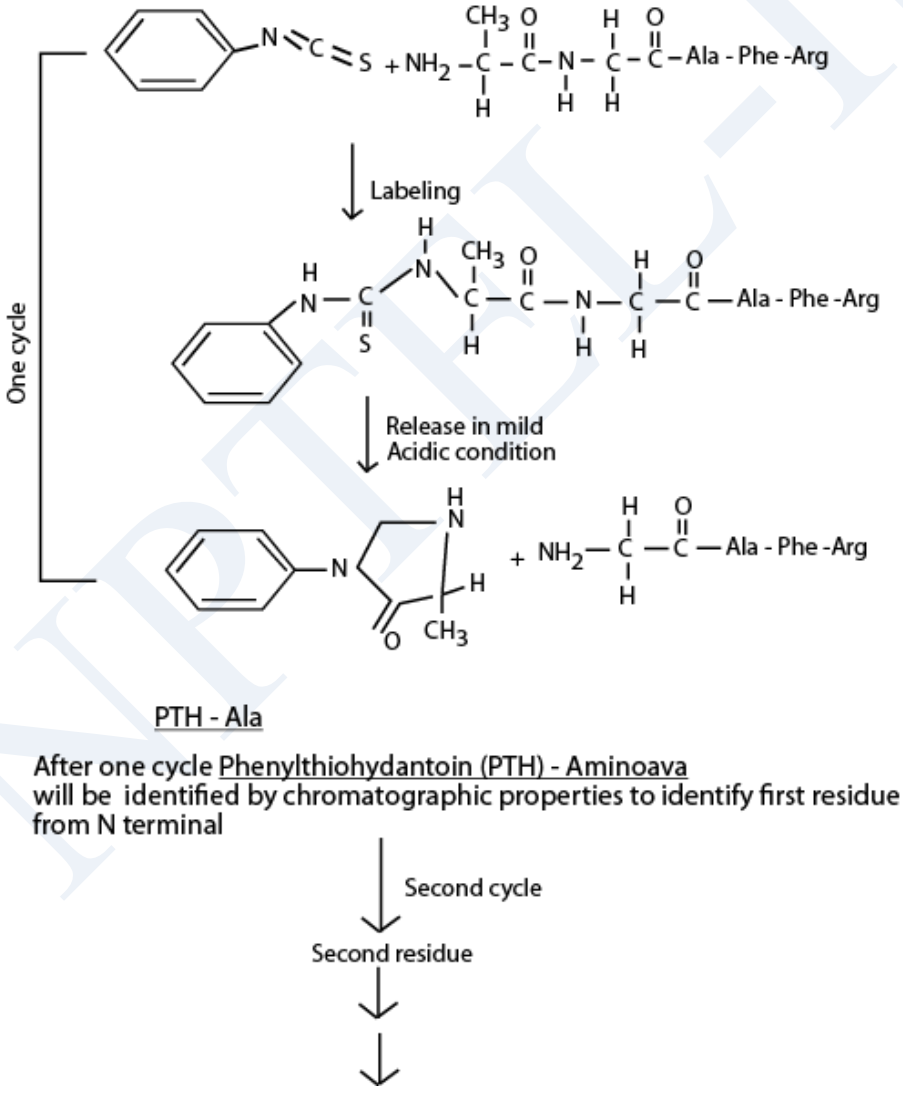


Figure 1: Scheme of protein sequencing by Edman degradation

Thus, reduction of disulfide bond in the polypeptide sequnce needed before sequencing process can be initiated. Reduction of free cysteine can be done by use of β-marcaptoethanol (Fig. 2)

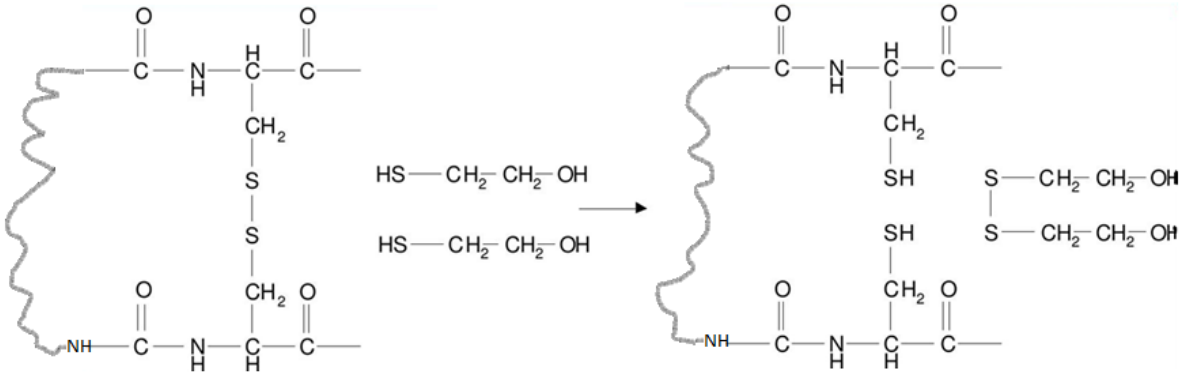


Figure 2: Chemical reaction showing reduction of disulfide bond by β-marcaptoethanol

As free cystein can re-oxidize to form disulfide it is necessary to block free cystein. This may be done by use of iodoacetic acid or acrylonitrile (free cysteine modification) as shown in Fig. 3.

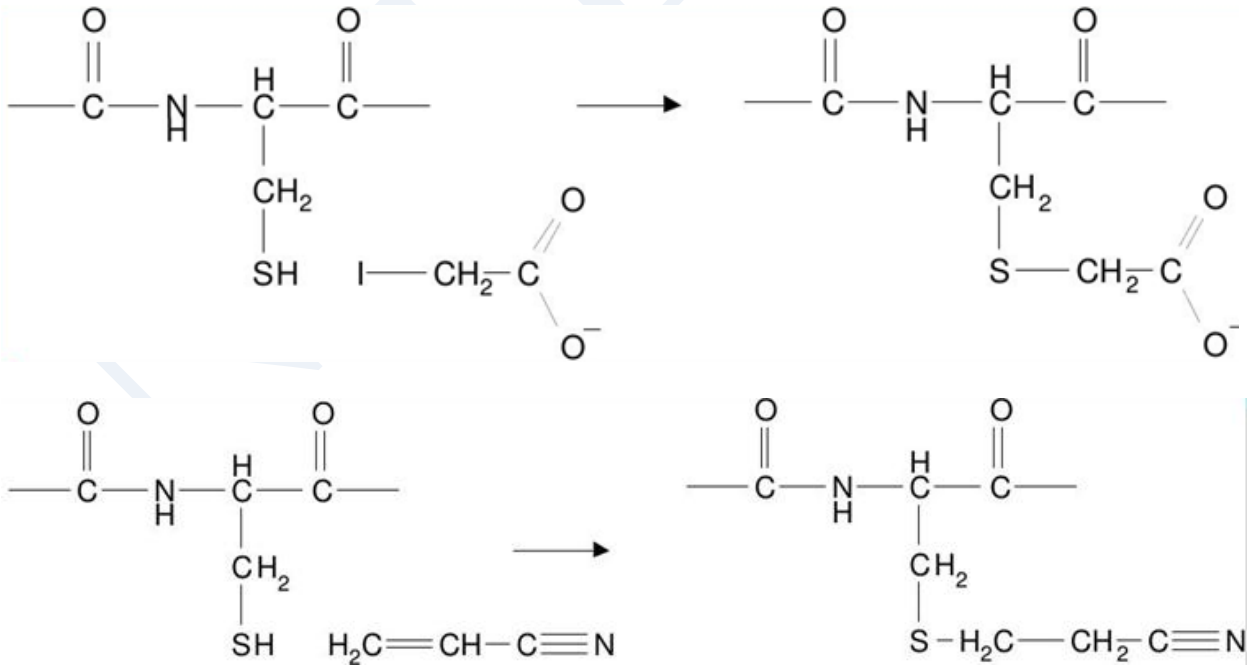


Figure 3: Free cystein may be blocked by use of iodoacetic acid or acrylonitrile

Other method for irreversible oxidation of disulfide bond is use of preformic acid. As shown in the figure below, performic acid oxidizes cysteine to negatively charge cysteic acid. Repulsion of negatively charged cysteic acid group prevents re-formation of disulfide and alkylation is not required. (Fig. 4)

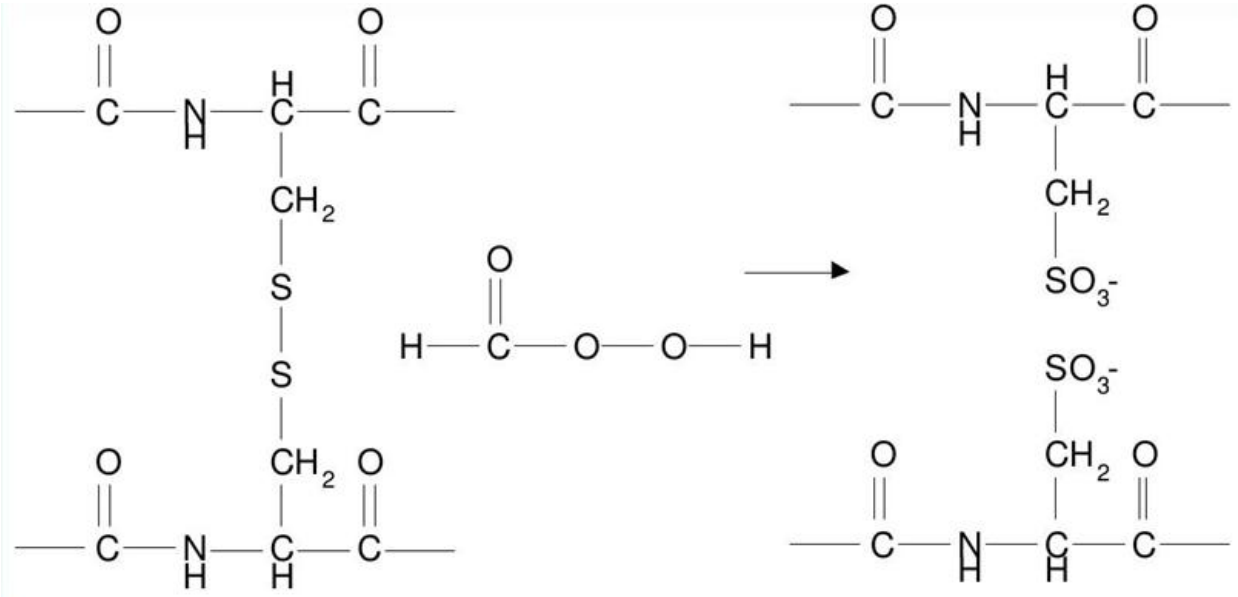


Figure 4: Irreversible oxidation of disulfide bond is use of preformic acid.

Further, the accuracy of each cycle is 98%. So after 60 steps the accuracy is less than 30%. Thus, this method cannot be used for sequencing of proteins larger than 50 amino acids. In case of larger proteins it has to be broken down to short peptide fragments using cleavage proteases such as trypsin (cleaves a protein at carboxyl side of lysine and arginine residues) or chymotrypsin (cleaves at carboxyl side of tyrosine, tryptophan and phenylalanine). Specific cleavage can also be achieved by chemical methods like cynogen bromide, which always cleaves at carboxyl side of methionine residue (a protein with 12 methionine will yield 13 fragment polypeptide on cleavage with cynogen bromide (CNBr).

Protein fragments after a protease (for example trypsine) will be separated and sequenced. Let us assume that the following two peptide sequences are obtained.

Gly-Gly -Phe -Ala -Lys

Thr- Phe- Gly-Lys

Now, a second method is used to cleave protein at other site. For example if we use chymotrypsin. Following sequence is found for one peptide sequence.

Gly-Lys-Gly-Gly-Phe

Now data from tyrosine cleavage fragment sequence and chymotrypsin cleavage fragment sequence can be analyze to get larger sequence information (Fig 5)

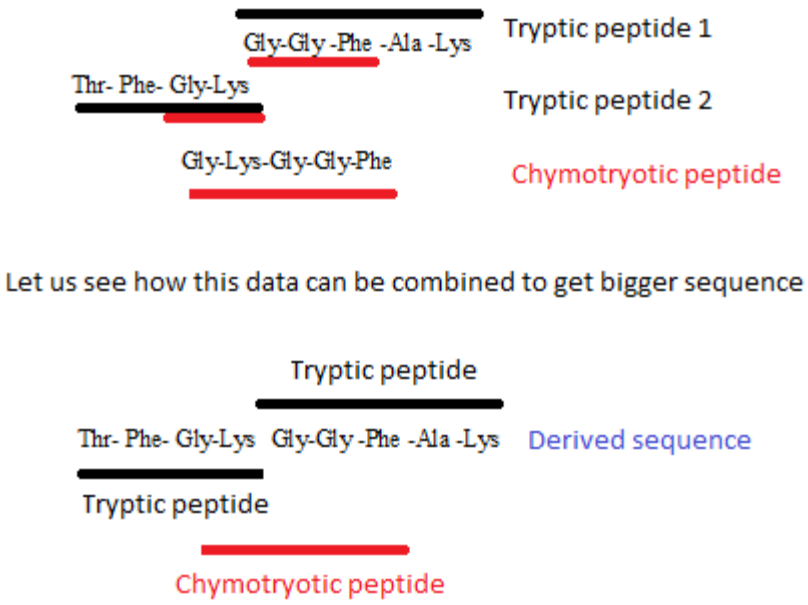
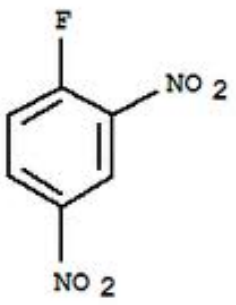


Figure 5: Sequences of smaller fragments may be overlapped and analyzed for complete sequence as explained above

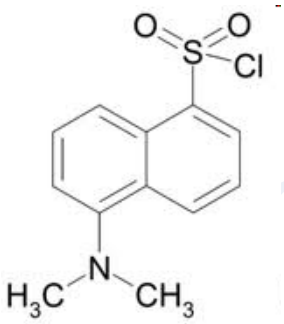
2) Protein sequencing using Sanger's reagent and dansyl chloride

Here, the N terminal amino acid of the protein is labeled by dyes like Sanger's reagent (fluoro-dinitrobenzene) or dansyl chloride. The labeled protein is then hydrolyzed by 6M HCl at 110 ºC by the above mentioned method and loaded in Dowex 50 column and the

elution profile is matched with the standard profile obtained from FNB or DNSCl derivative of all the amino acids, to obtain the N terminal amino acid. The reagents produce coloured derivatives which can be easily detected by absorbance (Fig. 6.)



Fluorodinitrobenzene



Dansyl Chloride

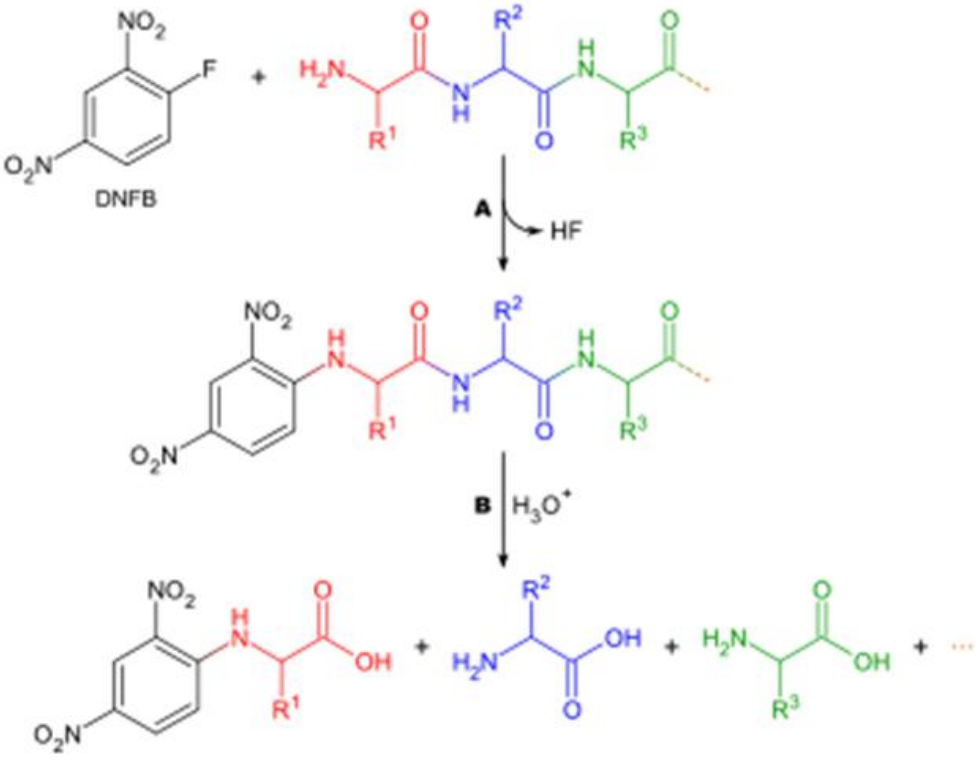


Figure source: Wikipedia

Figure 6: Protein sequencing using Sanger's reagent and dansyl chloride

Disadvantages of this method include:

- Once we get the N terminal amino acid, the protein is already hydrolyzed in constituent amino acids. Thus we cannot repeat the cycle with same sample. For second amino acid sequencing we require new stock of protein sample and the N-terminal residue need to be cleaved from the protein using an appropriate protease such as amino peptidase. This makes the process very tedious and complicated.
- These dyes selectively labels the amine groups present in the protein and therefore can label the amine groups present in the side chains as well, which may give erroneous results.

Protein sequencing using Molecular Biology techniques

If first few N-terminal amino acid of a protein is known, complete aminoacid sequence can be derived using Molecular Biology techniques. A simple example is as follow:

The genome sequence of *Calotropis procera*, a plant, or the sequence of procerain B, a novel cystein protease from the plant, gene is not yet known. Thus, the only information for cloning of cDNA we have is the fifteen N-terminal amino acid residues. The double stranded cDNA can be amplified with help of degenerate primer (based of N-terminal amino acid sequence) and oligo dT primer. Total RNA can be isolated from young leaf or latex of the plant and first strand of cDNA can be synthesised with oligo dT primer by reverse transcription. The second strand of cDNA can be synthesised and the subsiquent amplification of double stranded cDNA can be achived by PCR with degenerate primer as forward and oligo dT primer as reverse primmer. The amplified double stranded cDNA of expected size can be subjected to TA cloning and confirmed by sequencing. Once sequence of cDNA is availabe, it can be translated in protein sequnce.

[We shall study Few Molecular Biology techniques during coming lectures]