FEDERAL UNIVERSITY NDUFU-ALIKE IKWO
P.M.B. 1010, ABAKALIKI, EBONYI STATE
FACULTY OF SCIENCE AND TECHNOLOGY
DEPT. OF
BIOLOGY/MICROBIOLOGY/BIOTECHNOLOGY

BTG 405
(NUCLEOTIDE SEQUENCE ANALYSIS)

Course content:Nucleotide sequences assembly
Gene location and Identification
Protein sequence analysis and structure prediction

BY
ELEMBA, O. M.

Nucleotide Sequence assembly

- Nucleotide Sequence assembly refers to aligning and merging nucleotide fragments (reads i.e. short fragments) of a much longer DNA sequence in order to reconstruct the original sequence.

- It is the process of merging DNA fragments into larger "contigs" for subsequent analysis

- There is need for sequence assembly because DNA sequencing technology cannot read whole genomes at once, but rather reads small pieces of between 20 and 30000 bases, depending on the technology used.

- These reads (the short fragments) result from shotgun sequencing genomic DNA, or gene transcript (ESTs).

SEQUENCING APPROACHES

- The choice of assembly strategy depends on the sequencing method, and the choice of sequencing method may also depend on the organism that is being sequenced.

1. Sanger approach

2. Shotgun sequencing
   a. Whole Genome Shotgun (WGS)
   b. Hierarchical shotgun sequencing: (BAC-based or 'clone-by-clone')
3. Mixed strategy sequencing
4. EST sequencing
5. Massively parallel sequencing
6. Pyrosequencing
7. Sequencing-by-ligation
8. true Single Molecule Sequencing (tSMS)

COMPUTATIONAL ASSEMBLY

- Computational assembly is the only way to efficiently assemble sequenced fragments of DNA. However, a sufficient amount of high quality sequences are required.

- The assembly programs should be able to handle large data sets effectively and avoid misassemblies in the presence of large repetitive or duplicated regions and redundant sequences.

- To accomplish this, effective algorithms to handle large input data sets with the use of minimal computer time and memory are needed.

Two major assembling strategy is involve in computational assembly:

1. De novo assembling
2. Mapping assembling

- These two strategy can assemble different sequences irrespective of their source for example, sequences from the genome (genome assembler) and ESTs (EST assemblers)

1. De-novo assembling strategy: assembling short reads to create full-length (sometimes novel) sequences without matching with a template

- De-novo assemblies are orders of magnitude slower and more memory intensive than mapping assemblies.

- This is mostly due to the fact that the assembly algorithm needs to compare every read with every other read. E.g. the book shred.

2. Mapping assembling strategy: assembling reads against an existing backbone sequence, building a sequence that is similar but not necessarily identical to the backbone sequence.

- Compared to de novo assembly, the mapping of re-sequenced reads to a template genome is computationally easier.

- Efficient mapping tools are crucial and several tools for mapping of short reads are available.

- Most of the tools, *i.e.* MAQ, SOAP, SHRiMP or Eland, use seeding techniques that gain their speed from pre-computed hash look-up tables

BASIC PRINCIPLES OF ASSEMBLY

Majority assembly programs use the same basic scheme commonly known as the overlap-layout-consensus approach.

Essentially it consists of the following steps:

1. Sequence and quality data are read and the reads are cleaned.

2. Overlaps are detected between reads. False overlaps, duplicate reads, chimeric reads and reads with self-matches (including repetitive sequences) are also identified and left out for further treatment.

3. The reads are grouped to form a contig layout of the finished sequence.

4. A multiple sequence alignment of the reads is performed, and a consensus sequence is constructed for each contig layout (often along with a computed quality value for each base).

5. Possible sites of misassembly are identified by combining manual inspection with quality value validation.

Prior to the assembly, the electropherogram for a given sequence is interpreted as a sequence of bases (a read) with associated quality values, these values reflect the log-odds score of the bases being correct.

The basecaller PHRED is often used, however alternatives exist, *e.g.* the CATS basecaller. The reads can then be screened for any contaminant DNA such as *Escherichia coli*, cloning or sequencing vector. Low quality regions can be identified and removed.

Base quality values can be used in computation of significant overlaps and in construction of the multiple alignments.

Issues that can affect the final assembly (other than the obvious quality of sequence data) are:

- The size of the inserts, whether the sequencing was uni- or bi-directional,
- The library construction,
- The cloning vector,
- The selection of clones to be sequenced, and
- The availability of additional information (consensus genome, ESTs, known verified genes, gene maps, etc.).

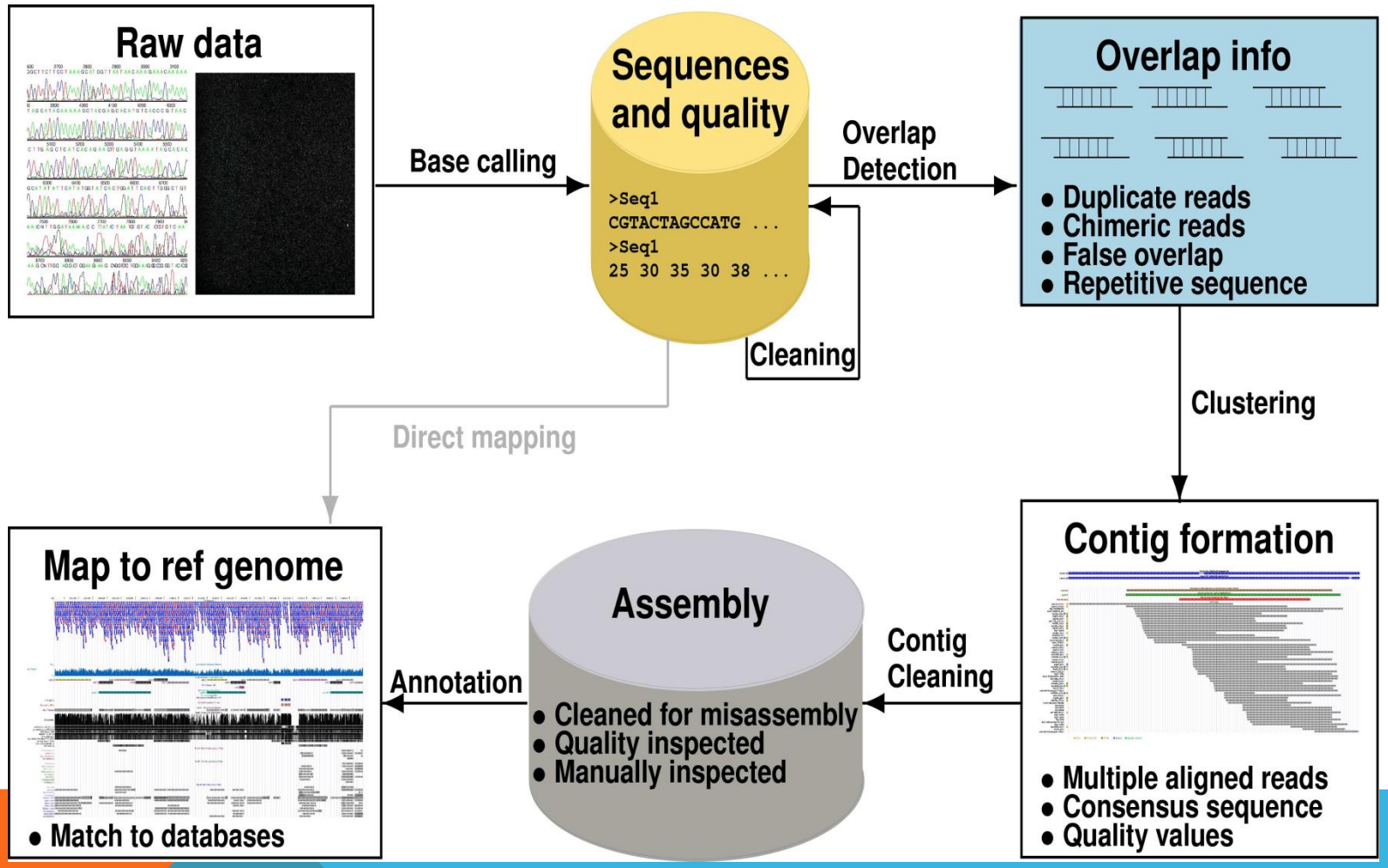


Figure 1: Schematic diagram showing a typical sequence assembly

Finishing

- When an assembly has been completed, specific parts of the assembly usually need to be re-examined, perhaps due to low quality of the data, low (or no) coverage of the sequence, sites under suspicion of mis-assembly, etc.

- The re-examination are usually dealt with in an elaborate process

- Analysis of the assembled contigs can be performed with a number of tools.
- One is Consed, which allows navigation of the assembled contigs and reads, problematic regions can be searched for with different criteria, and regions can be tagged for further inspection.

- Others are Autofinish, BACcardi and GAP4 all of which has different strengths and purposes.

General Assembler differences

When different assemblers try to put reads together they essentially work from the same input, but the assemblers differ in the way they utilize the sequence information, and in the way this is combined with additional information.

In general the differences fall in the following categories.

1. Overlaps: A lot of different methods are used to find potential overlaps between sequences. Some are based on BLAST (*e.g.* geneDistiller), while other assemblers use various other methods to find similarities between reads.

2. Additional information: Depending on how the sequence reads are produced, some additional information might be available. This information might consist of read pair information, BAC clone information, base quality information, etc. Some assemblers use this data to impose additional structure on the assembly of the sequences (*e.g.* GigAssembler).

3. Short read assembly: De novo assembly of the micro reads generated from next generation sequencing platforms is still challenging. While assemblers have been developed and applied to assemble bacterial genomes successfully on larger genomes the assembly is performed by mapping the micro reads to reference genomes. The major next generation sequencing platforms all have built-in software to handle this task, *eg.* GS Reference mapper, Gerald for Solexa, SOLiD systems etc.

In SOLiD systems the mapping tool “mapreads” converts reference sequences into color space and perform the mapping in color space.

Assemblers

- Different assemblers that have been created over time and are used. An overview with shorts presentations of the different assemblers are given on the web-page http://genome.ku.dk/resources/assembly/methods.html.

A. Assembler for small DNA sequences

- One of the (relatively) early assemblers is PHRAP , which is still in use, both in itself (for small DNA sequence sets),

B. WGS assemblers

- RePS
- Phusion,
- JAZZ and ATLAS
- Celera assembler,
- CAP3
- RAMEN
- PCAP
- TIGR assembler
- STROLL, and ARACHNE2

C. WGS and ESTs assemblers

- Some new approaches to assembly have been attempted, among them are mira and TRAP which try novel ways to deal with repetitive sequences by checking the trace and quality files.

- An emerging approach is to use more explicit graph based programs, such as: Euler, Partial ordered alignment (POA), Velvet, Splicing graphs, Asmodeler and xtract, where the last three are used specifically for ESTs.

D. ESTs assemblers

Other programs that analyze ESTs are :

- Splicing graphs
- Asmodeler
- Xtract
- TGICL
- StackPack
- PaCE
- HiddenMarkovModel (HMM) Sampling and
- geneDistiller
- PAVE assembler (Program for Assembling and Viewing ESTs)

Programs are used in the scaffolding stage

- Finally, some programs are used in the scaffolding stage, where contigs are processed and put in order, e.g. GigAssembler and Bambus

BTG 405
(COMPUTER NUCLEOTIDE SEQUENCE ANALYSIS)

Course content:Protein sequence analysis and structure prediction

BY

ELEMBA, O. M.

PROTEINS

Introduction

- Proteins or polypeptides play an outstanding part in all cell activities.

- They are linear chains consisting of a sequence of 20 amino acids in different combinations linked exclusively by peptide bonds.

- The peptide bonds are formed by the reaction of the primary amino group of one amino acid with primary carboxyl group of another amino acid with the elimination of a molecule of water (Figure 1). Thus, it is a condensation reaction.

- This type of linkage causes a polarity for the polypeptide chain.

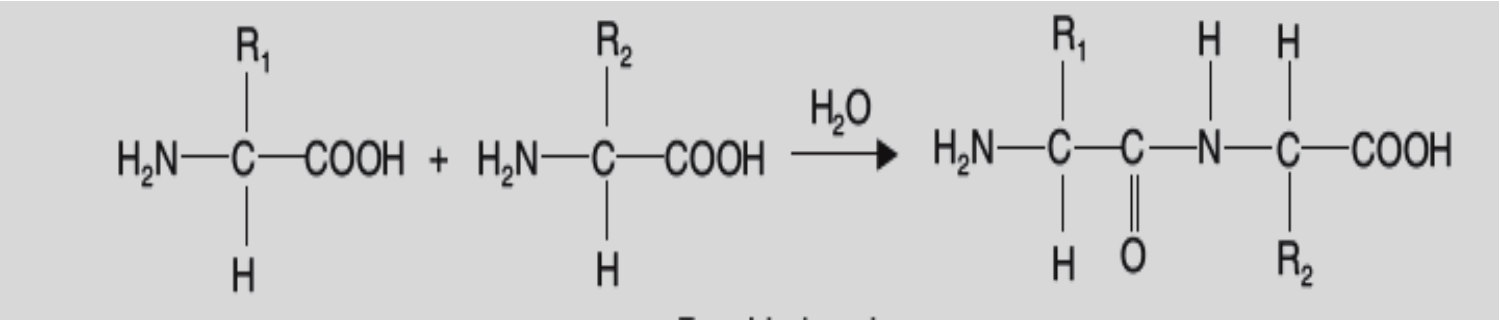


Figure 1: formation of peptide bond

- One end has the amino group and is called the N-terminus, while the other end is terminated by a free carboxyl group and is called the C-terminus.

- Amino-acid sequences are written from N- to C-terminus, the direction in which protein synthesis proceeds.

- The exact sequence of amino acids (also called the protein's primary structure) is determined by the nucleotide sequence of the gene, the part of the DNA strand, which codes for the protein.

- The three-dimensional structure and the function of the protein are very closely dependent on the amino-acid sequence of the polypeptide.

- The different combinations and sequences of amino acids are responsible for the diverse nature and functions of proteins (Table 1).

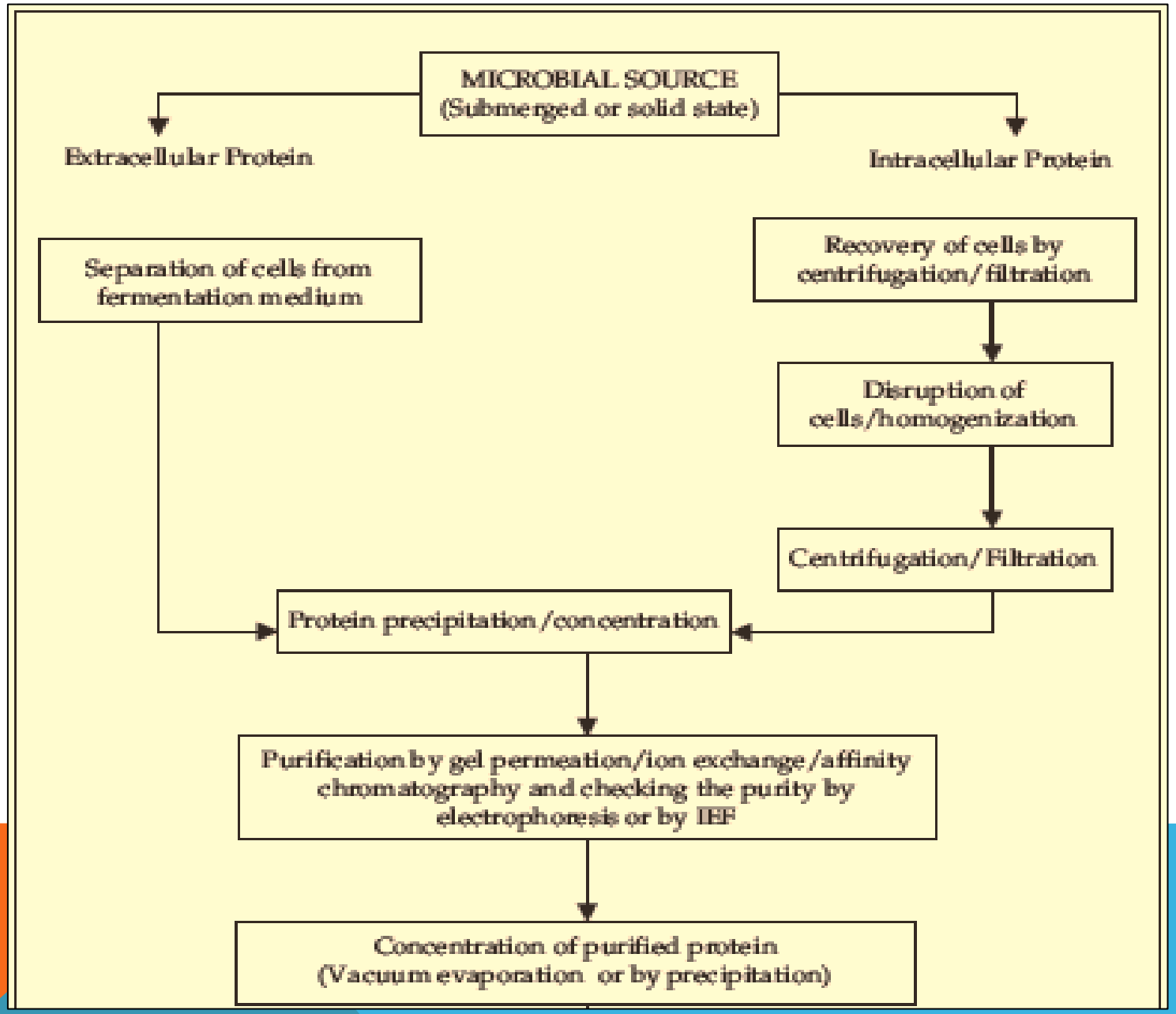
- They act as biological catalysts (= enzymes), take part in the regulation of the cell's metabolism and in the interaction between cells, and are required for the generation of specific structures.

TABLE 1: SOME IMPORTANT PROTEINS AND THEIR FUNCTIONS

Protein	Site of location	Function
Collagen	Connective tissue	Gives tensile strength
Thrombin	Blood	Blood coating
Trypsin	Pancreatic juice	Cleaves the polypeptide during protein digestion
Amylase	Salivary secretion	Starch hydrolysis
Insulin	Secreted by islet cells of pancreas into blood stream	Controls the blood sugar level
Immuniglobulins	Blood	Immunity
Endorphins	Brain	Regulates the brain
Rhodopsin	Retinal cells of eye	activities
Cytochrome	Mitochondrial membranes	Vision Cellular respiration

TYPES OF PROTEIN

- There are three types of proteins based on their complexity.
- Some proteins are made up of a single polypeptide chain and are known as simple proteins.
- Those proteins having two or more polypeptide chains are called complex proteins.
- In some cases, the protein molecule is associated with a non-protein component known as the prosthetic group. Such proteins are known as conjugated proteins.
- The non-protein component may be metallic ions such as Zn+ in the case of carbonic anhydrase, hem part of hemoglobin enclosing Fe+ ion in it, or organic molecules such as vitamin derivatives such as NAD and NADP, nucleotides such as ATP and GTP, or maybe sugars, oligosaccharides, or various types of lipids.


MICROBIAL SOURCE
(Submerged or solid state)
Extracellular Protein
Intracellular Protein
Separation of cells from fermentation medium
Recovery of cells by centrifugation/ filtration
Disruption of cells/homogenization
Centrifugation/ Filtration
Protein precipitation /concentration
Purification by gel permeation/ion exchange /affinity chromatography and checking the purity by electrophoresis or by IEF
Concentration of purified protein
(Vacuum evaporation or by precipitation)

ANALYSIS OF PROTEIN: AMINO ACID COMPOSITION AND PROTEIN-SEQUENCING

Amino acid Composition

- To determine the amino acid composition the peptide is first hydrolyzed into its constituent amino acids by heating in 6 N HCl at 110°C for 24 hours.

- The amino acids in the hydrolysate can be separated by ion-exchange chromatography and hydrolyzed by reacting them with ninhydrin.

- Alpha amino acids treated this way give an intense blue color, whereas amino acids, such as proline, give a yellow color.

- The concentration of amino acids in a solution is proportional to the optical absorbance of the solution after heating it with ninhydrin.
- This technique can detect a microgram (10 n mol or nanograms) of an amino acid.
- After getting the information about the amino acid composition and relative quantity of each amino acid,
- one can proceed to do the sequencing of amino acids for a particular protein or polypeptide.

Amino acid Sequencing

- The amino acid sequence of a protein is very important because it is essential to know the structure and function of that protein;
- it can also help in identifying and isolating the gene code for the protein.
- So obtaining at least a partial amino acid sequence is a critical first step in studying many proteins.

There are two basic procedure for sequencing a protein

1. Sangers method
2. Edman degradation method

The sanger's method

- This involve the use of certain reagent such as fluoro-dinitro-benzene (FDNB) known as Sanger's reagent
- This reagent react specifically with the free NH_2 group of the amino acid at the N-terminal of a polypeptide.
- This yields a yellow DNP derivative of the amino acid on acid hydrolysis
- This derivative which can be separated and identified by ion-exchange chromatography.
-

- This procedure cannot be used on the same sample of polypeptide because the peptide is totally hydrolyzed on the acid hydrolysis step.

- But Sanger managed to use this technique with new samples of peptide in each cycle of experiments to sequence insulin.

- He used more than a gram of insulin for completing this task. Now this method is used only for identifying the N-terminal amino acid of a polypeptide and not for sequencing because of the above-mentioned demerit.

Edman Degradation Procedure

- The most popular direct protein-sequencing technique in use today is the Edman degradation procedure.

- The Edman reaction is a series of chemical reactions, which remove one amino acid at a time from amino terminus of a protein, releasing an amino-acid derivative, phenylthiohydantoin (PTH), that may be chromatographically identified (reversed phase).

- The release of amino acids from the amino terminus in a sequential manner is possible because the Edman procedure consists of three chemical reactions, which proceed under different pH conditions

- The first step is the coupling step , which occurs at high pH values and results in formation of phenylthiocarbamoylated (PTC) amino groups on the protein.

- This happens when the the Edman reagent, Phenylthiocyanate, reacts with the NH_2 group of the terminal amino acid and forms an intermediate, phenylthiocarbomyle derivative.

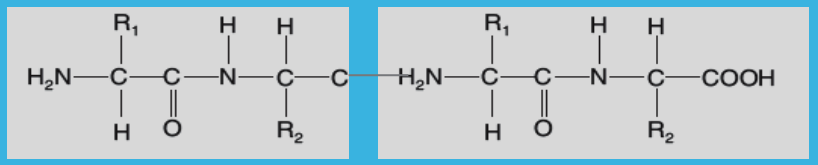
- The second reaction is the cleavage step, which occurs at low pH, resulting in release of an anilionothiazolinone (ATZ) form of the amino acid and regeneration of a free amino terminus on the protein.

- Thirdly: The ATZ-amino acid is converted to the phenylthiohydantoin (PTH) derivative in a separate reaction, generally exposure to strong acid.

- The PTH-amino acid can be separated and identified by HPLC (High Performance Liquid Chromatography). This leaves the intact peptide short of one amino acid.
- The Edman procedure can be repeated on the shortened peptide obtained in the previous cycle for identifying the second amino acid of the polypeptide from the N-terminal.
- The whole method, including the reaction steps and the identification of PTH amino acid derivative, is automated and the instrument that can carry out these reactions is known as a sequenator.

- Using this automated instrumentation, more than 100 amino acids of a polypeptide can be sequenced efficiently.

- Polypeptides and proteins of high molecular weight have to be fragmented into smaller polypeptides of 50 to 100 amino acids before carrying out the sequencing by a sequenator



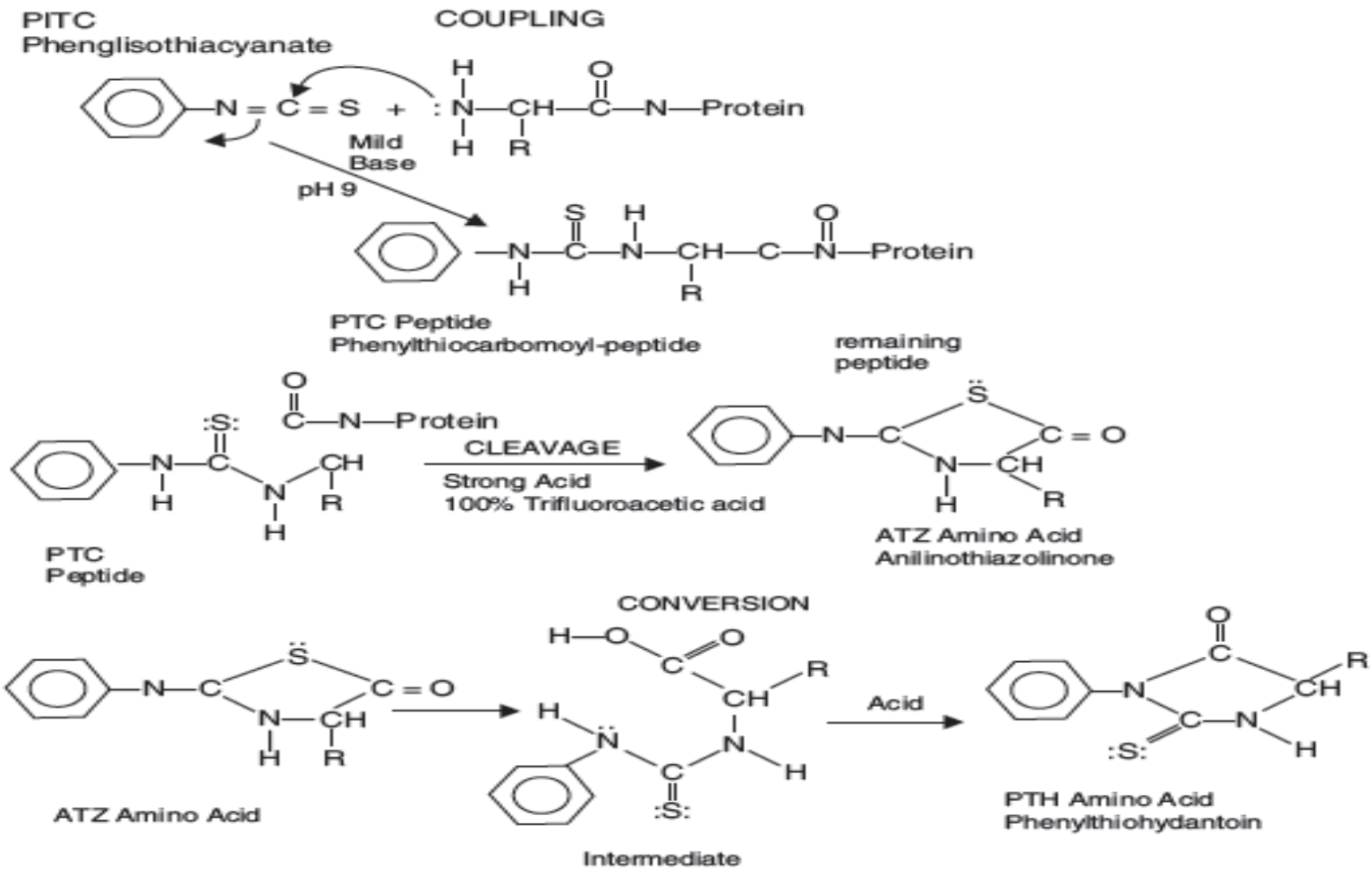


Figure 2: Different steps of Edman degradation reaction as operated in a sequenator.

Protein-sequencing Strategies

- Most proteins are large in size and contain large numbers of amino acid residues. Therefore, it is necessary to fragment these long polypeptide chains into small fragments, which can be sequenced completely by the Edman degradation reaction.
- Fragmentation of polypeptides into smaller bits can be carried out either by chemical methods or by enzymatic cleavage. Both chemical and enzymatic cleavage are very specific, and the cleavage profile of a polypeptide by a specific chemical or enzyme can be used for the identification of an unknown protein.
- Proteases or certain chemical reagents are used to selectively cleave some of the peptide bonds of a protein (Table 2).
- The smaller peptides fragments formed are then isolated and subjected to sequencing by the Edman degradation procedure.
- The chemical reagent, cyanogen bromide (CNBr) reacts specifically with methionine residues to produce peptides with C-terminal homoserine lactone residues and new N-terminal residues. Since most proteins contain very few methionine residues, treatment with CNBr usually produces only a few peptide fragments. Reaction of CNBr with a polypeptide chain containing three internal methionine residues should generate four peptide fragments. Each fragment can then be sequenced from its N-terminus.
- In the final stage of sequence determination, the amino acid sequence of the original large polypeptide chain can be deduced by lining up the amino acid sequence of cleaved peptide fragments and matching the overlapping sequences of peptide fragments (Figure 2)

Table 2: specific cleavage of polypeptides

Reagents	Cleavage site
Chemical Cleavage	
Cyanogenbromide	Carboxyl side of methionine residues
O-Idosobenzoate	Carboxyl side of tryptophan residues
Hydroxylamine	Asparagine–glycine bonds
2-Nitro-5-thiocyanobenzoate	Amino side of cysteine residues
Enzymatic Cleavage	
Trypsin	Carboxyl side of lysine and arginine residues
Chymotrypsin	Carboxyl side of tyrosine, tryptophan, phenylalanine, leucine, and methionine
Clostripain	Carboxyl side of arginine residues
Staphylococcal protease	Carboxyl side of glutamate and aspartate
Thrombin	Carboxyl side of arginine residues
Carboxypeptidase A	Amino side of C-terminal amino acids except arginine, lysine, and proline

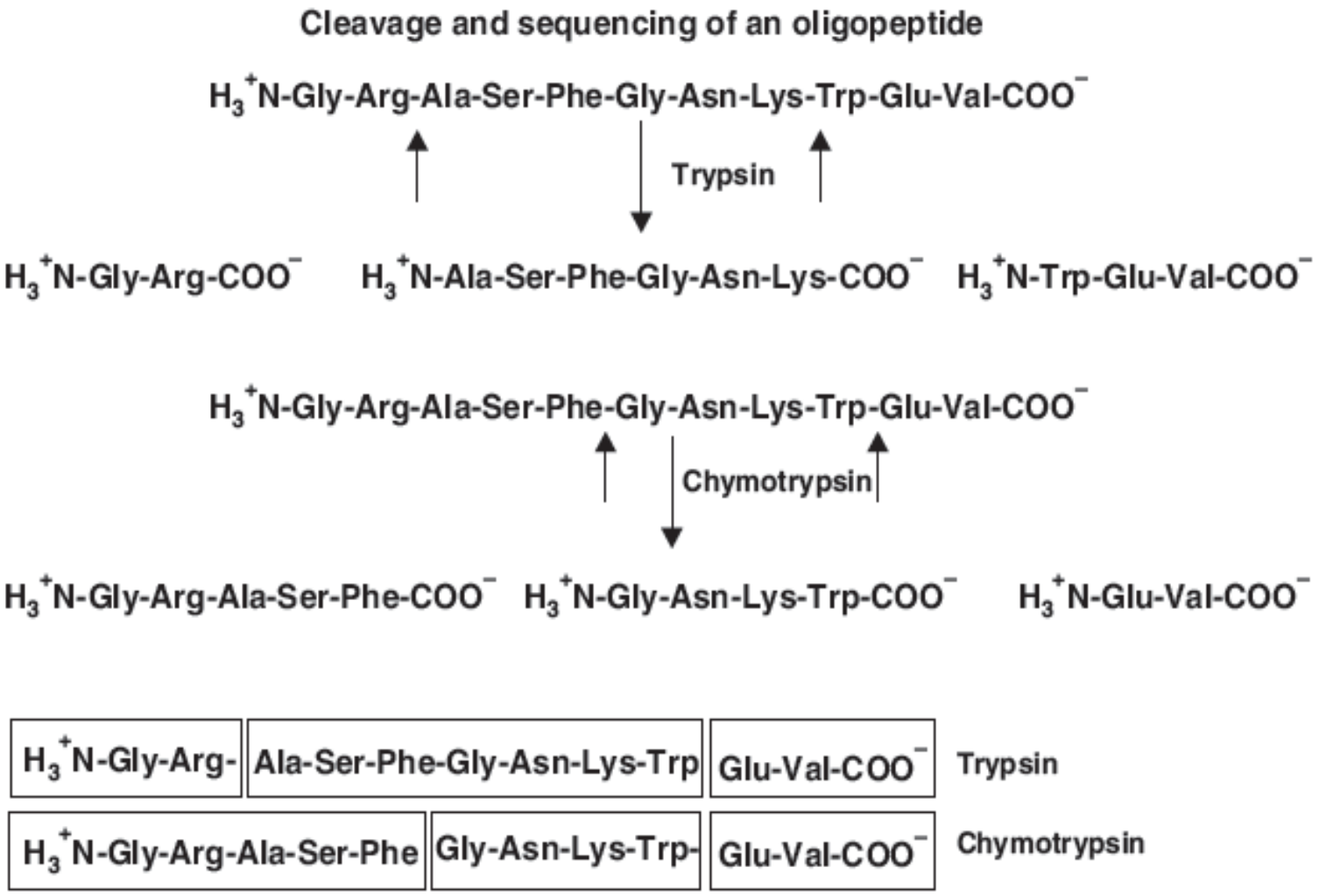


Figure 2: Schematic representation of cleavage, sequencing and alignment of an oligopepetide

STRUCTURE IDENTIFICATION

- Identification of the structure of protein is important because, the function of a protein depends on the three dimensional structure adopted by folding the linear chain of amino acids after translation in the environment of the cell.

- There are a number of physical techniques like:

- X-ray crystallography

- Mass spectrometry (NMR: nuclear magnetic resonance spectrometry), and

- Dual polarisation interferometry techniques are used in protein analysis and structural elucidation.

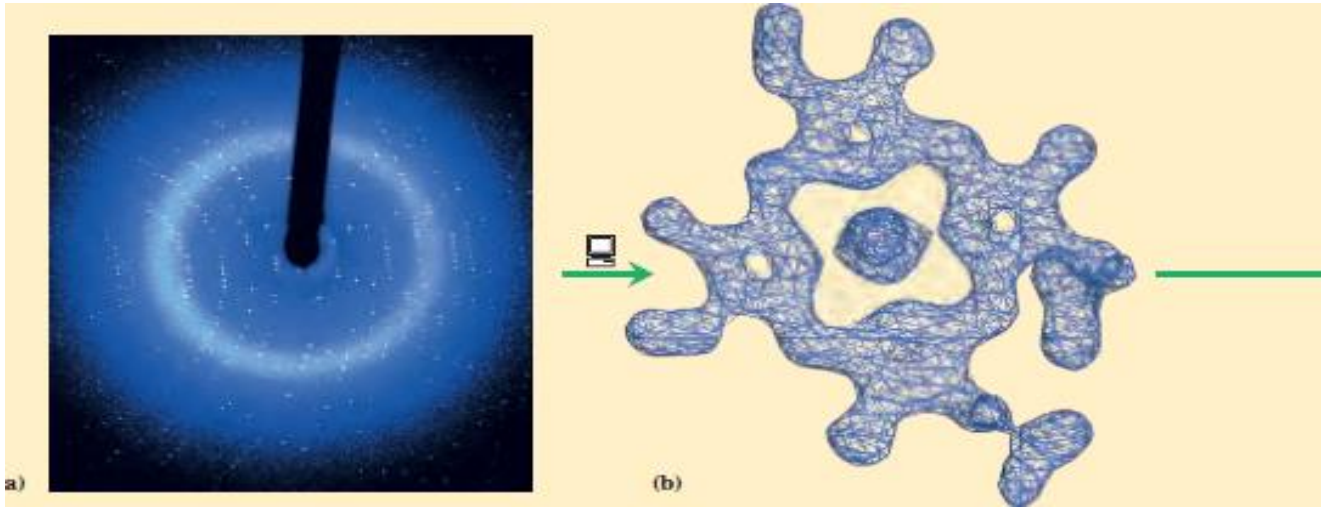
X-ray crystallography method

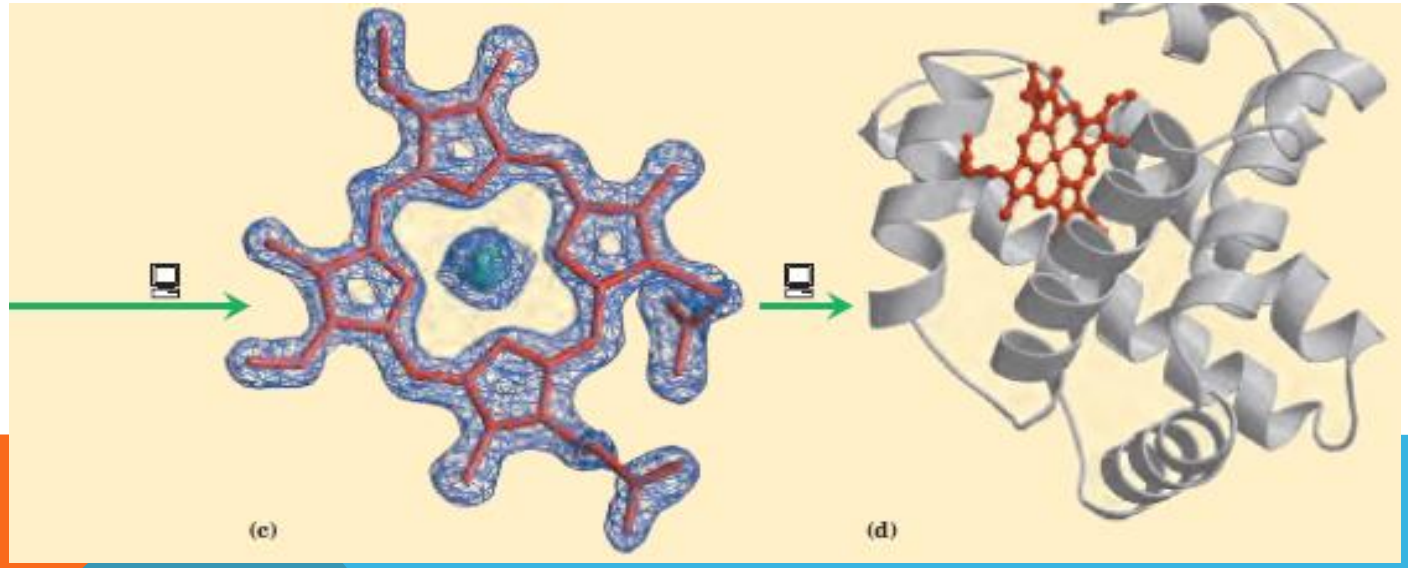
- X-ray crystallography is a method of determining the arrangement of atoms within a crystals.
- when a beam of X-rays strikes a crystal of protein, it causes the beam of light to spread into many specific directions.
- from the angles and intensities of these diffracted beams scientist can produce a three-dimensional picture of the density of electrons within the crystal.
- From this electron density, the mean position of the atoms in the crystal can be determined, as well as their chemical bonds, their disorder and various other information.

Operationally, there are several steps in x-ray structural analysis.

- **Once a crystal is obtained, it is placed in an x-ray beam between the x-ray source and a detector, and a regular array of spots called reflections is generated.**

- **The spots are created by the diffracted x-ray beam, and each atom in a molecule makes a contribution to each spot.**
- **An electron-density map of the protein is reconstructed from the overall diffraction pattern of spots by using a mathematical technique called a Fourier transform. In effect, the computer acts as a "computational lens."**
- **A model for the structure is then built that is consistent with the electron-density map.**


a)
(b)


(c)
(d)

Nuclear magnetic resonance (NMR)

- Certain proteins do not readily form crystals.
- Nuclear magnetic resonance (NMR) is used to determine the structures of these type of peptides in solution.

- The basic technique is that some atoms, such as natural isotopes of nitrogen behave like small magnets and can switch between spin states in applied magnetic field. This happens by the absorbance of low wavelength electromagnetic radiations and thus generate NMR spectra.

- The resonance frequency for each atom is unique and is usually influenced by the surrounding electron density. This implies that changes(usually referred to as chemical shift) allows the distinction between, for instance, an aryl and an aromatic group. The decay of the magnetic resonance in a molecule also depends on the structure and spatial arrangement within a molecule.

- The analysis of NMR spectra thus allow the building a distance constraints which could be used to determine the three dimensional spatial arrangement of an atom.