UNIT 5

PROTEIN AND GENE MANIPULATION

CHAPTER 1

RECOMBINANT DNA TECHNOLOGY

5.1.1. Introduction

Every day the newspapers tell you about the remarkable feats of science and technology in helping man combat disease and improve his environment. Many of these involve the use of a powerful technique called gene cloning or Recombinant DNA Technology (RDT). Using such technology bacteria in the past were engineered to produce human insulin a hormone which fights diabetes, yeast cells were made to produce Hepatitis B vaccine, plants such as cotton were made insect resistant (Bt-cotton) and even as you read this chapter projects are on to engineer bacteria to cleanup environmental waste such as polythene. Have you not wondered how Scientists are able to achieve all this? The present chapter will introduce you to the main techniques used in gene cloning along with some important applications for you to understand and marvel at the simplicity and power of this area of biotechnology.

The basic steps involved in RDT are illustrated schematically below **Fig. 1** :

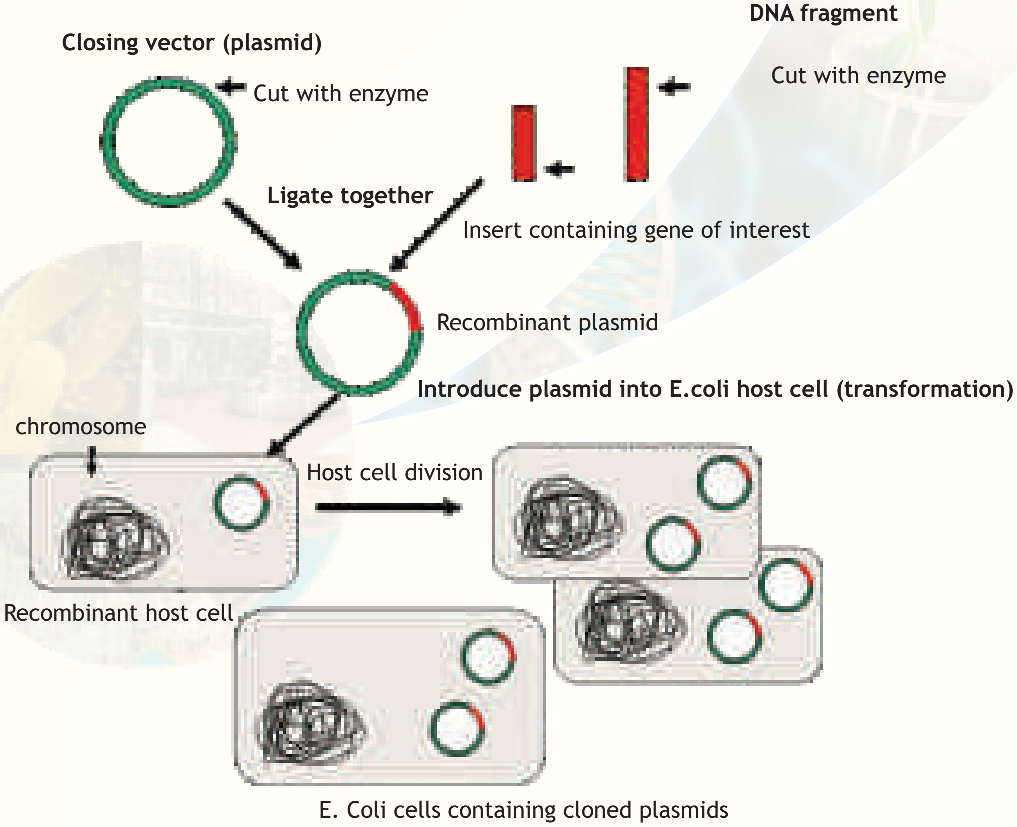


Fig. 1. Schematic representation of the basic steps in RDT.

The steps involved are:

1. Isolation of a DNA fragment containing a gene of interest that needs to be cloned (called as **insert**).

2. Generation of a recombinant DNA (rDNA) molecule by insertion of the DNA fragment into a carrier DNA molecule called **vector** (e.g. plasmid) that can self replicate within a host cell.

3. Transfer of the rDNA into an *E. coli* host cell (process called **transformation**).

4. Selection of only those host cells carrying the rDNA and allowing them to multiply thereby multiplying the rDNA molecules.

The whole process thus can generate either a large amount of rDNA (**gene cloning**) or a large amount of protein expressed by the insert. The first rDNA molecules to be generated using these procedures were established by the combined efforts in 1973 by the molecular biologists Paul Berg, Herbert Boyer, Annie Chang and Stanley Cohen.

5.1.2. Tools of rDNA technology

In order to generate recombinant DNA molecules, we not only require the vector and insert DNA but also a method to precisely cut these DNA molecules and then join them together **(ligation)**. Several molecular tools are required to perform the various steps and these are described in the following sections.

Restriction Enzymes: The Molecular Scissors

The foundations of rDNA technology were laid by the discovery of restriction enzymes. These enzymes exist in many bacteria where they function as a part of a defence mechanism called the Restriction-Modification System. This System consists of two components:

1. A restriction enzyme that selectively recognises a specific DNA sequence and digests any DNA fragment containing that sequence. The term restriction is derived from the ability of these enzymes to restrict the propagation of foreign DNA (*e.g.* Viral/phage DNA) in a bacterium.

2. A modification enzyme that adds a methyl group to one or two bases within the sequence recognised by the enzyme. Once a base is modified by methylation, the sequence cannot be digested. It is thus obvious that the Restriction-Modification enzyme system within a given bacterium protects its DNA from digestion by methylation but can digest foreign DNA which is not protected by similar methylation.

Different species of bacteria contain their own sets of restriction endonucleases and corresponding methylases. Three main classes of restriction endonucleases- type I, type II and type III are present, of which, only type II restriction enzymes are used in rDNA technology as they recognise and cut DNA within a specifc sequence typically consisting of 4-8 bp. This sequence is referred to as a restriction site and is generally palindromic, which means that the sequence in both DNA strands at this site read same 5' to 3' direction. Type II restriction enzymes are named after the bacterial species they have been isolated from. For example a commonly used restriction enzyme *Eco*RI isolated from the bacterial species *E. coli* is named so with the first three italicised alphabets referring to the genus (E) and species (co), the capital R referring to the strain (RY 13) and the number designated with the roman numeral (I) indicating that it was the first enzyme to be isolated from this strain of bacteria. Restriction enzymes were first discovered and studied by the molecular biologists W. Arber, H. Smith and D. Nathans for which they were awarded the Nobel Prize in 1978. Today more than a thousand restriction enzymes are available for genetic engineers to use. Some commonly used restriction enzymes (type II) along with their source and restriction- modification sequences have been listed in **Table1** below.

Table 1. Type II restriction enzymes, their sources, recognition and cleavage sites.

Restriction enzyme	Microbial source	Recognition sequence
Alu I	*Arthrobacter luteus*	5'A-G↓-C-T 3’ 3'T-C↑-G-A 5’
*Bam*HI	*Bacillus amyloliquefaciens*	5'G↓-G-A-T-C-C 3' 3'C-C-T-A-G↑-G 5'
*Eco*RI	*Escherichia coli*	5'G↓-A-A-T-T-C 3' 3'C-T-T-A-A↑-G 5'
*Eco*RII	*Escherichia coli*	5'↓C-C-T-G-G 3' 3'G-G-A-C↑-C 5’
*Hae*III	*Haemophilus aegyptus*	5'G-G↓-C-C 3’ 3'C-C↑-G-G 5’
*Hind*III	*Haemophilus influenza*	5'A↓-A-G-C-T-T 3’ 3'T-T-C-G-A↑-A 5’
*Pst*I	*Providencia stuartii*	5'C-T-G-C↓-A-G 3’ 3'G↑-A-C-G-T-C 5’
*Sal*I	*Streptomyces albus*	5'G↓-T-C-G-A-C 3’ 3'C-A-G-C-T↑-G 5'

The exact kind of cleavage produced by a restriction enzyme is important in the design of a gene cloning experiment. Some cleave both strands of DNA through the centre resulting in a blunt or flush end. These are also known as symmetrical cuts. From **Table I** it is obvious that the enzyme *AluI* cuts symmetrically.

$$\begin{array}{l} 5'\text{-A-G}\overset{\downarrow}{\text{-}}\text{C-T-}3' \\ 3'\text{-T-C}\underset{\uparrow}{\text{-}}\text{G-A-}5' \end{array} \longrightarrow \begin{array}{l} 5'\text{-A-G} \\ 3'\text{-T-C} \end{array} + \begin{array}{r} \text{C-T-}3' \\ \text{G-A-}5' \end{array}$$

However *Eco*RI cuts in a way producing protruding and recessed ends known as sticky or cohesive ends because these ends can base pair and stick the DNA molecule back together again. Such cuts are termed staggered.

$$\begin{array}{l} 5'\text{-G}\overset{\downarrow}{\text{-}}\text{A-A-T-T-C-}3' \\ 3'\text{-C-T-T-A-A}\underset{\uparrow}{\text{-}}\text{G-}5' \end{array} \longrightarrow \begin{array}{l} 5'\text{-G-}3' \\ 3'\text{-C-T-T-A-A-}5' \end{array} + \begin{array}{r} 5'\text{-A-A-T-T-C-}3' \\ 3'\text{-G-}5' \end{array}$$

Note that *Eco*RI generates 5' overhangs at the cut site (and 3' recessed ends). Hence if two different DNA fragments containing *Eco*RI recognition sites are cleaved and mixed the sticky ends can bind and generate a hybrid or recombinant DNA **(Fig. 2)**

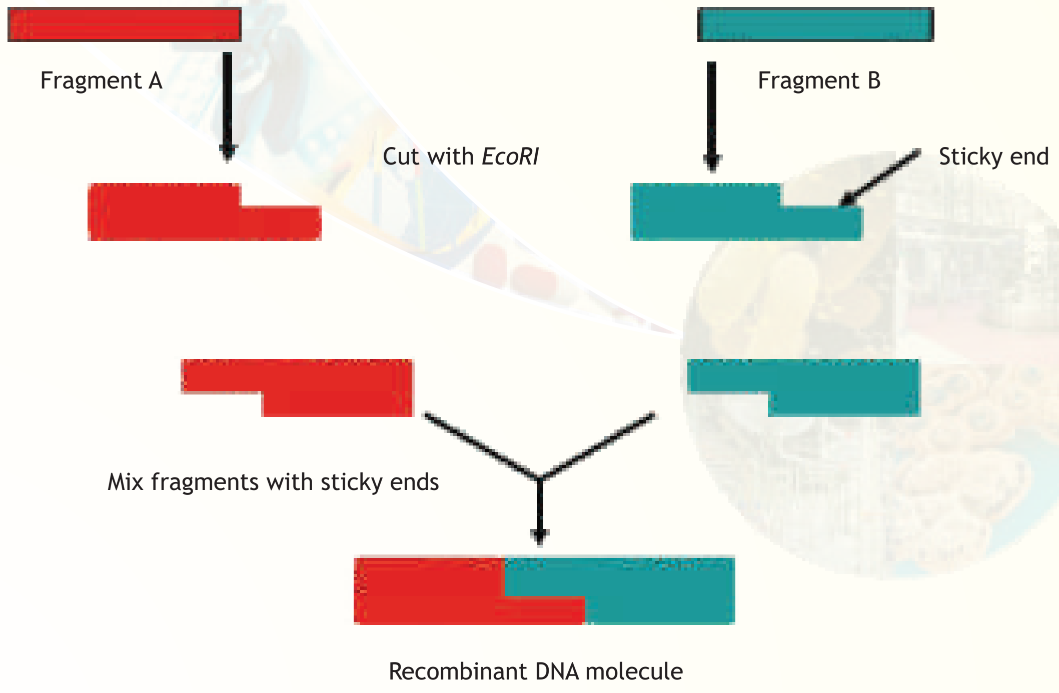


Fig. 2. Construction of rDNA using fragments from different sources.

Restriction Fragment Length Polymorphism (RFLP)

The DNA isolated from an individual organism has a unique sequence and even the members within a species differ in some parts of their sequence. The restriction sites would also vary and hence if DNA from a given individual was subjected to digestion with a restriction enzyme the fragments generated would vary when compared with another individual's DNA similarly digested. This variation in size (length) of the restriction enzyme generated fragments among individuals within a given species is termed RFLP. A schematic representation of how RFLPs are generated is given below **(Fig. 3)**. A major application of this technique is DNA fingerprinting analysis which is explained below.

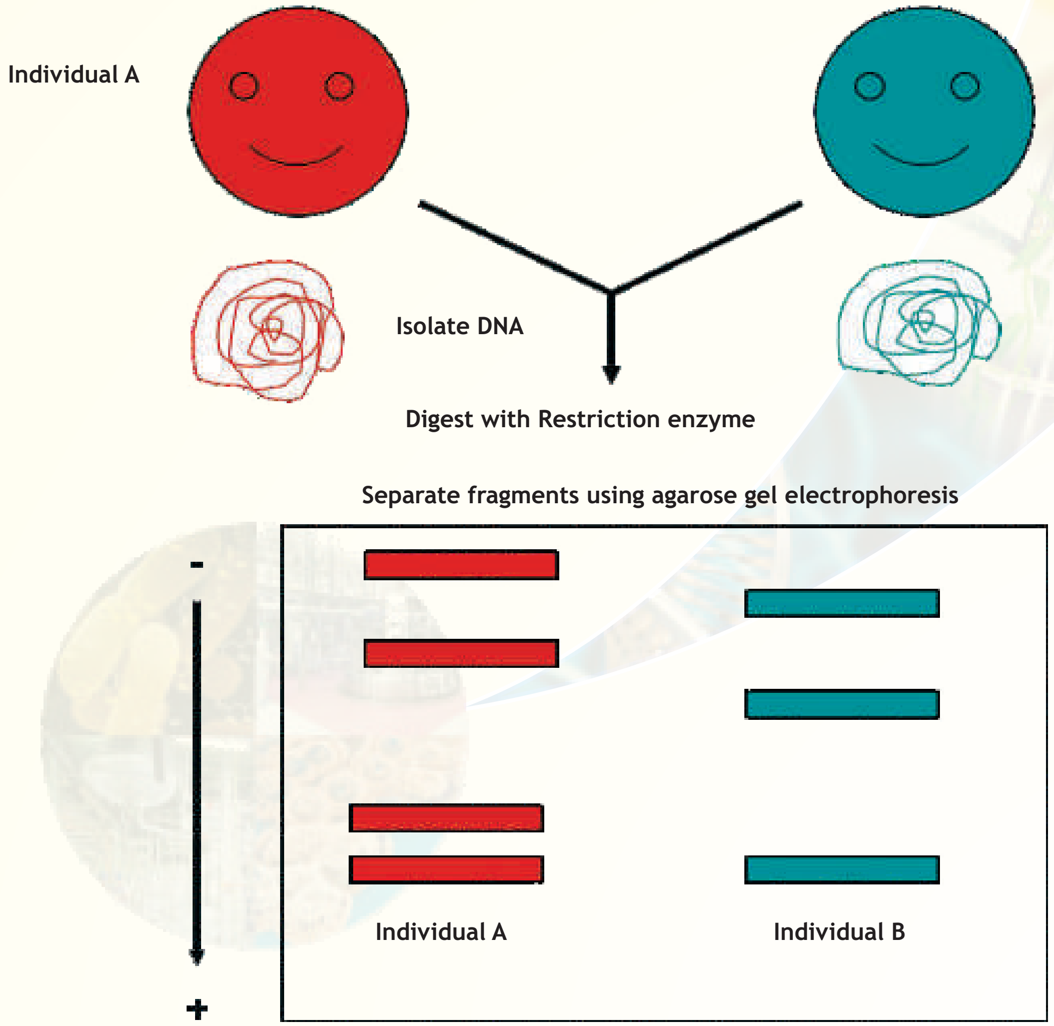

Individual A
Individual B
Isolate DNA
Digest with Restriction enzyme
Separate fragments using agarose gel electrophoresis
-
+
Individual A
Individual B

Fig. 3. RFLP technique.

Individuals except identical twins vary in their RFLP pattern as indicated schematically in the agarose gel electrophoresis. Hence the term DNA fingerprint is used and this is the basis of a major technique used in forensic science to identify and relate individuals.

Other enzymes used in cloning

In addition to restriction enzymes, there are several other enzymes that play an important role in rDNA technology. Two of these are DNA ligase and alkaline phosphatase.

DNA ligase: This enzyme forms phosphodiester bonds between adjacent nucleotides and covalently links two fragments of DNA . The reaction requires one of the fragments to have a 5' phosphate residue and the other a 3' hydroxyl group. In a previous section it was indicated how two fragments cut with *Eco*RI could stick together; DNA ligase seals this by forming a covalent bond. DNA ligase isolated from the bacteriophage T4 is frequently used to ligate different DNA fragments in order to generate rDNA molecules.

Alkaline phosphatase: Ligation requires the presence of a 5'phosphate group. If some of the fragments are treated with alkaline phosphatase to remove their phosphate groups then these cannot ligate within themselves and are forced to ligate with other fragments containing 5'phosphate groups. Hence this is a useful strategy to prevent self-ligation which would otherwise lead to wasteful ligation of fragments treated with restriction enzymes. An insert is ligated to the vector in generating rDNA as the vector is prevented from self-ligation by treating it with alkaline phosphatase. Alkaline phosphatase used for this purpose is purified from bacteria or calf intestines.

Vectors: Vehicles for cloning

Another major component of a gene cloning experiment is a vector such as a plasmid. A vector serves as a vehicle to carry a foreign DNA sequence into a host cell. A vector must possess certain features:

1. It must contain an origin of replication (ori) so that it is independently able to replicate within the host. This implies that any foreign insert it carries is automatically replicated.

2. It should incorporate a selectable marker, a gene whose product can identify the host cells containing the vector. Selectable markers include genes conferring antibiotic resistance, enzymes such as β-galactosidase which can turn substrates blue in the vicinity of the host cell colony and gene expressing Green Fluorescent Protein (GFP) which cause host cells containing the vector to fluoresce when viewed under UV light.

3. The vector must also have one unique restriction enzyme recognition site which can be used for cutting and introducing an insert. Most of the commonly used cloning vectors have more than one restriction site, they contain a Multiple Cloning Site (MCS) or polylinker. The MCS provides flexibility in the choice and use of restriction enzymes.

4. Another desirable feature of a cloning vector is that it should be small in size thereby facilitating entry/transfer into a host cell.

   A number of vectors have been developed incorporating these features but only plasmids and bacteriophage based vectors will be discussed.

Plasmids

Plasmids are extrachromosomal, self-replicating, usually circular, double-stranded DNA molecules found naturally in many bacteria and also in some yeasts. Although plasmids are not essential for normal cell growth and division, they often confer useful properties to the host such as resistance to antibiotics that can be a selective advantage under certain conditions. The plasmid molecules can be present as 1 or 2 copies or in multiple copies (500-700) inside the host organism. These naturally occurring plasmids have been modified to serve as vectors in the laboratory and are by far the most widely used, versatile and easily manipulated vectors.

One of the earliest plasmid vectors to be constructed was pBR322 **(Fig. 4a)**. This plasmid contains two different antibiotic resistance genes and recognition sites for several restriction enzymes. A popular series of plasmid cloning vectors is the pUC family **(Fig. 4b)**. These vectors contain a region of the *lacZ* gene that codes for the enzyme β-galactosidase. This region also contains a polylinker and thus insertion of a foreign DNA into any of the restriction sites will result in an altered non-functional enzyme. During screening of recombinant plasmid containing host cells the absence of β-galactosidase activity is indicative of plasmids containing the insert.

The plasmid vectors described above can replicate only in *E. coli*. Many of the vectors used in eucaryotic cells are constructed such that they can exist both in the eukaryotic cell and *E. coli*. Such vectors are known as shuttle vectors. These vectors contain two types of origin of replication and selectable marker genes, one set which functions in the eukaryotic cells (*e.g.* yeast) and another in *E. coli*. An example of a shuttle vector is the yeast plasmid Yep **(Fig. 4c)**. In the case of plants, a naturally occurring plasmid of the bacterium *Agrobacterium tumefaciens* called Ti plasmid has been suitably modified to function as a vector.

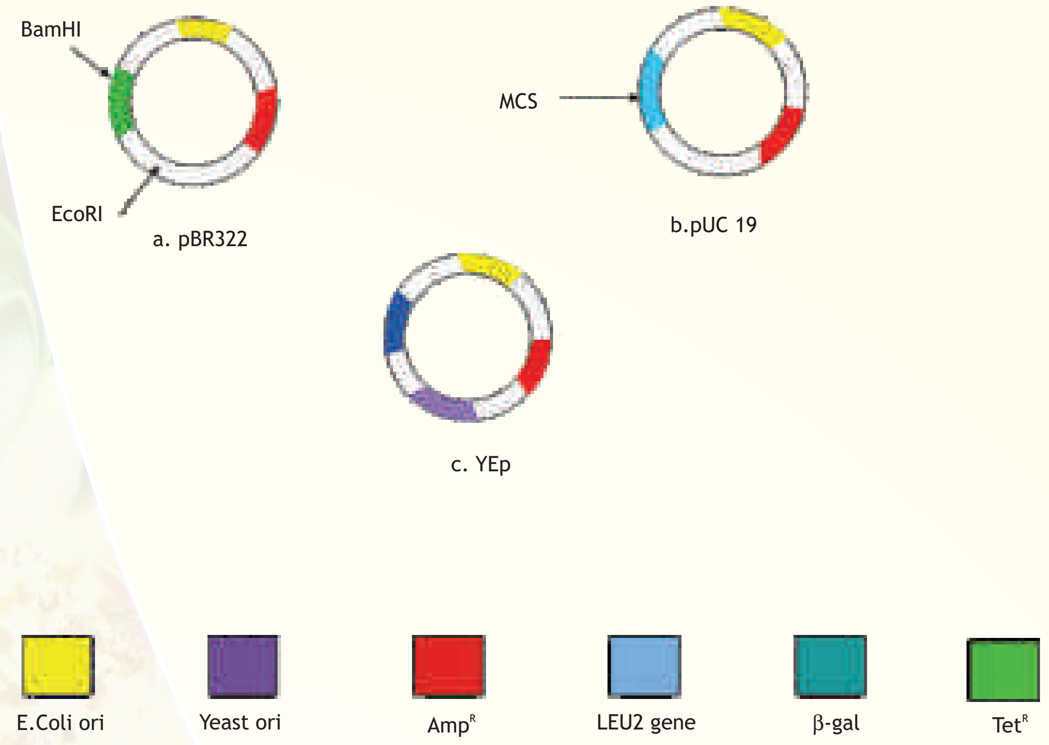


Fig. 4. Schematic representation of some plasmids (not drawn to scale).

The *LEU2* gene (see Yep) codes for an enzyme which is needed for the synthesis of the amino acid leucine. Yeast cells having this plasmid can grow on a medium lacking leucine and hence can be selected over cells not containing the plasmid.

The discussion so far has focussed on cloning a DNA fragment to sufficiently amplify it. However the goal of the cloning experiment may be to produce a foreign protein in the host. In such cases the fragment containing the gene expressing the foreign protein is incorporated into the vector along with signals necessary for transcription and translation in the given host. Vectors which are suitable for expressing foreign protein are called expression vectors and one such vector is the pUC 19.

Vectors based on bacteriophages

Bacteriophages are viruses that infect bacterial cells by injecting their DNA into them and consequently take over the machinery of the bacterial cells to multiply themselves. The injected DNA hence is selectively replicated and expressed in the host bacterial cell resulting in a number of phages which eventually extrude out of the cell ***(lytic pathway)*** and infect neighbouring cells. This ability to transfer DNA from the phage genome to specific bacterial hosts during the process of viral infection gave scientists the idea that specifically designed phage based vectors would be useful tools for gene cloning experiments. Two phages that have been extensively modified for the development of cloning vectors are lambda (λ) and M13 phages.

Bacteriophage lambda has a double stranded, linear DNA genome containing 48,514 bp, in which 12 bases on each end are unpaired but complementary. These ends therefore are sticky and are called cohesive or cos sites and are important for packaging DNA into phage heads. An important feature of the lambda genome is that a large fragment in the central region of its genome is not essential for lytic infection of *E. coli* cells. Therefore, vectors have been designed such that this region can be replaced by foreign insert DNA. These phage based vectors allow cloning of DNA fragments up to 23 kb in size.

M13 is a filamentous phage which infects *E. coli* having a pilus (protrusion) which is selectively present in cells containing a F plasmid (called F+ cells). The genome of the M13 phage is a single stranded, circular DNA of 6407 bp. Foreign DNA can be inserted into it without disrupting any of the essential genes. In the life cycle of the phage following infection of the host *E.coli* cell the single stranded DNA is converted to a double-stranded molecule which is referred to as the Replicative Form (RF). The RF replicates until there are about 100 copies in the cell. At this point DNA replication becomes asymmetric and single stranded copies of the genome are produced and extruded from the cell packaged with protein as M13 particles **(Fig. 5)**. The major advantages of developing vectors based on M13 are that its genome is less than 10 kb in size; the RF can be purified and manipulated exactly like a plasmid. In addition, genes cloned into M13 based vectors can be obtained in the form of single stranded DNA. Single stranded forms of cloned DNA are useful for use in various techniques including DNA sequencing and site-directed mutagenesis, techniques which will be discussed in a latter section.

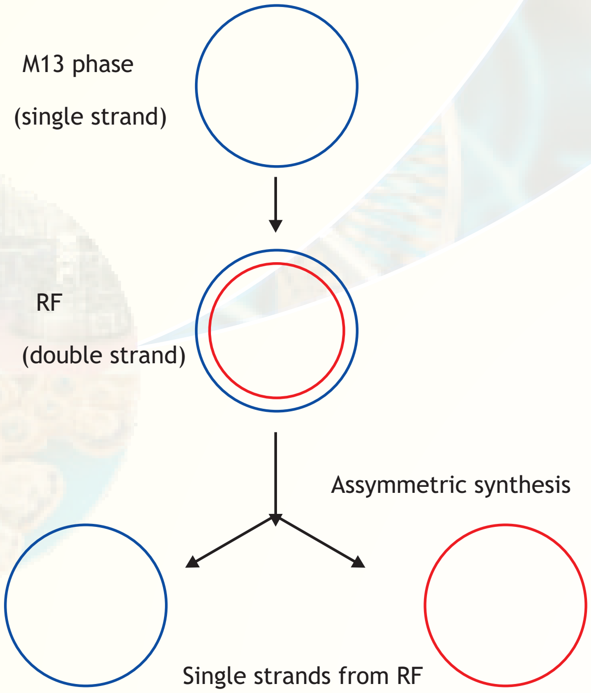


Fig. 5. Life cycle of M13 phage.

Cosmids

Cosmids have been constructed by combining certain features of plasmid and the 'cos' sites of phage lambda. The simplest cosmid vector contains a plasmid, origin of replication, a selectable marker, suitable restriction enzyme sites and the lambda cos site. Cosmids can be used to clone DNA fragments up to 45 kbp in size.

YAC vectors

YACs or Yeast Artificial Chromosomes are used as vectors to clone DNA fragments of more than 1 Mb in size. Therefore, they are useful in cloning larger DNA fragments as required in mapping genomes such as in the Human Genome Project. These vectors contain a *teleomeric* sequence, the centromere and an autonomously replicating sequence, features required to replicate linear chromosomes in yeast cells. These vectors also contain suitable restriction sites to clone foreign DNA as well as genes to be used as selectable markers.

BAC vectors

BACs or Bacterial Artificial Chromosomes are vectors based on the natural, extrachromosomal plasmid from *E. coli* - the fertility or F plasmid. A BAC vector contains genes for replication and maintenance of the F plasmid, a selectable marker and cloning sites. These vectors can accommodate inserts up to 500 kb and are used in genome sequencing projects. **Table 2** lists the common cloning vectors with the size of insert that can be cloned into them.

Table 2. Common cloning vectors.

Vector Type	**Insert size (kb)**
Plasmid	0.5-8
Bacteriophage lambda	9-23
Cosmid	30-40
BAC	50-500
YAC	250-1000

Animal and Plant viral vectors

You have already learnt how bacteriophages have been used to derive suitable vectors for gene cloning experiments in *E. coli*. Similarly, viruses that infect plant and animal cells have also been manipulated to introduce foreign genes into plant and animal cells. The natural ability of viruses to adsorb to cells, introduce their DNA and replicate, have made them ideal vehicles to transfer foreign DNA into eukaryotic cells in culture. A vector based on Simian Virus 40 (SV40) was used in the first cloning experiment involving mammalian cells. A number of vectors based on other type of viruses like Adenoviruses and Papillomavirus have been used to clone genes in mammals. At present, retroviral vectors are popular for cloning genes in mammalian cells. In case of plants, viruses like Cauliflower Mosaic Virus, Tobacco Mosaic Virus and Gemini viruses have been used with limited success.

Host Cells

The tools described in the previous sections will result in the generation of recombinant DNA molecules in the laboratory. Eventually the propagation of these DNA molecules must occur inside a living system or host. Many types of host cells including *E. coli*, yeast, animal and plant cells are available for gene cloning and the type of host cell used depends on the aim of the cloning experiment. *E. coli* has become the most widely used organism in rDNA technology because its genetic make-up has been intensively studied, it is easy to handle and grow, can accept a range of vectors and has been extensively studied for safety. Another major advantage of using *E. coli* as host cells is that under optimal conditions the cells divide every 20 minutes making it possible to clone large amounts of foreign DNA and if the appropriate signals are incorporated into the vector large amounts of recombinant proteins are available for therapeutics and other uses.

For the expression of eukaryotic proteins, eukaryotic cells are often preferred because to be functionally active, proteins require proper folding and post translational modifications such as glycosylation which is not possible in prokaryotic *(E. coli)* cells. Even cloned eukaryotic genes containing introns cannot be processed in *E. coli* thereby necessitating the use of only eukaryotic host cells. Yeast cells have been used extensively for functional expression of eukaryotic genes because of several features. Yeasts are the simplest eukaryotic organisms (unicellular) and like *E. coli* have been extensively characterised genetically, easy to grow and manipulate and large amounts of cloned genes or recombinant proteins can be obtained from yeast cultures grown in fermentors (large culture vessels). Plant and animal cells may also be used as hosts in rDNA experiments and cells can be grown in tissue culture or can be induced and manipulated to form whole organisms (creation of transgenic animals and plants).

5.1.3. Making rDNA

The first step in the construction of a recombinant DNA molecule is to isolate the vector and the fragment containing the gene to be cloned. The vector and target DNA fragment are separately digested with the same restriction enzyme such as *Eco*RI which generates sticky ends. The vector is then treated with alkaline phosphatase enzyme so that later in the ligation step the vector does not self ligate. The cut vector and DNA fragment are mixed in a suitable ratio and then ligated with the enzyme DNA ligase to yield a recombinant vector containing insert. This procedure is schematically explained in **Fig. 6**.

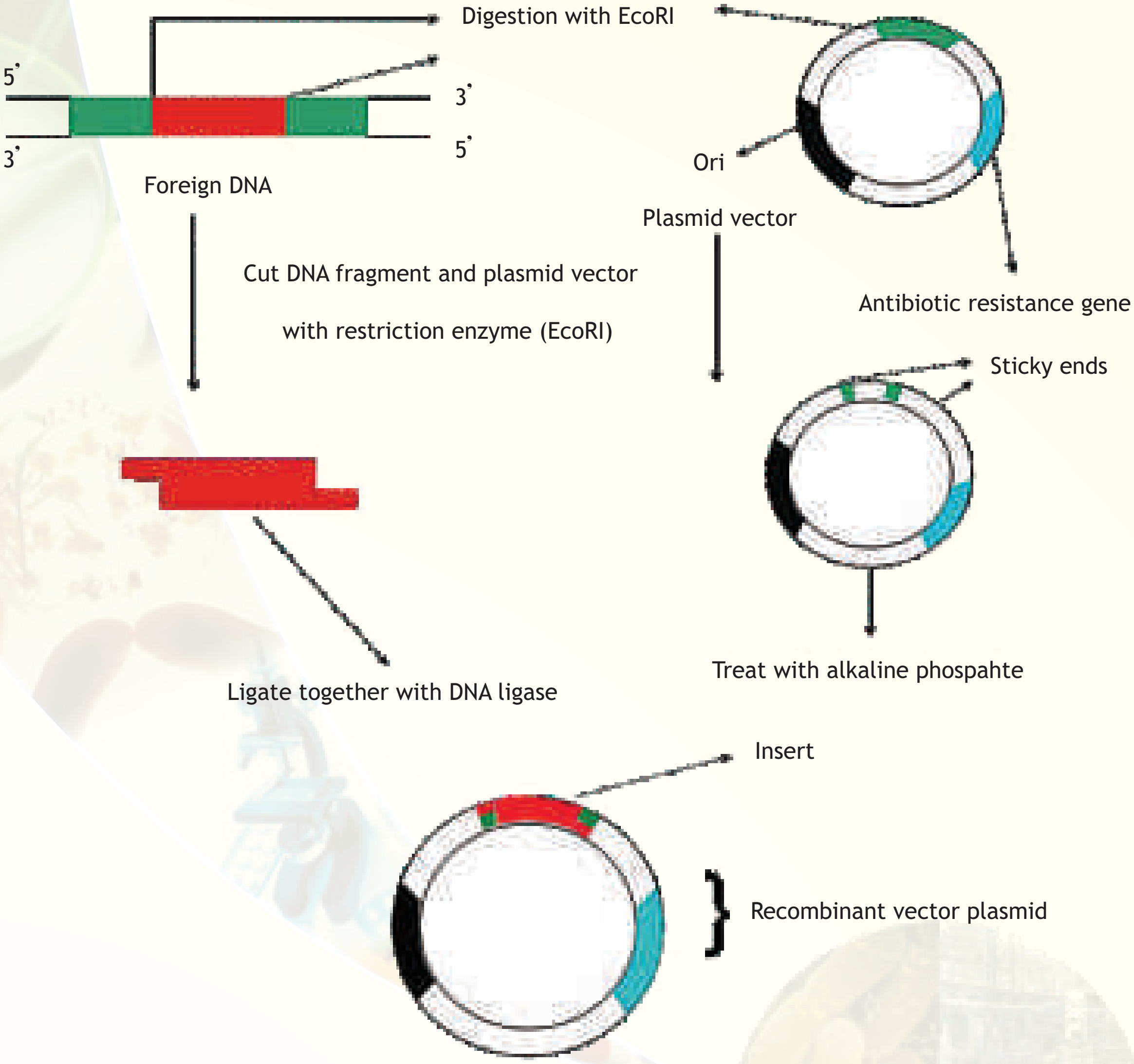

Digestion with EcoRI
5'
3'
3'
5'
Foreign DNA
Ori
Plasmid vector
Cut DNA fragment and plasmid vector
with restriction enzyme (EcoRI)
Antibiotic resistance gene
Sticky ends
Treat with alkaline phospahte
Ligate together with DNA ligase
Insert
Recombinant vector plasmid

Fig. 6. Making recombinant plasmid

5.1.4. Introduction of rDNA into host cells

The next step after a recombinant molecule has been generated is to introduce it into a suitable host. There are many methods to introduce recombinant vectors and these are dependant on several factors such as the vector type and host cell. Some commonly used procedures are discussed below.

Transformation: In rDNA technology, the most common method to introduce rDNA into living cells is called transformation. In this procedure, bacterial cells take up DNA from the surrounding environment. Many host cell organisms such as *E. coli*, yeast and mammalian cells do not readily take up foreign DNA and have to be chemically treated to become competent to do so. In 1970, Mandel and Higa found that *E. coli* cells become markedly competent to take up external DNA when suspended briefly in cold calcium chloride solution.

Transfection: Another method to transfer rDNA into host cells involves mixing the foreign DNA with charged substances like calcium phosphate, cationic liposomes or DEAE dextran and overlaying on recipient host cells. Host cells take up the DNA in a process called transfection.

Electroporation: An electric current is used to create transient microscopic pores in the recipient host cell membrane allowing rDNA to enter.

Microinjection: Exogenous DNA can also be introduced directly into animal and plant cells without the use of eukaryotic vectors. In the procedure of microinjection, foreign DNA is directly injected into recipient cells using a fine microsyringe under a phase contrast microscope to aid vision.

Biolistics: A remarkable method that has been developed to introduce foreign DNA into mainly plant cells is by using a gene or particle gun. Microscopic particles of gold or tungsten are coated with the DNA of interest and bombarded onto cells with a device much like a particle gun. Hence the term biolistics is used.

Another method of introducing foreign genes is by the natural genetic engineer *Agrobacterium tumefaciens*. This method and its principle will be covered in the chapter of plant cell culture in Unit VI.

5.1.5. Identification of Recombinants

Once a recombinant DNA molecule has been introduced into appropriate host cells, it becomes imperative to select only those cells which have the rDNA from those of the original host cells which have not taken up the DNA. All procedures described in the previous section are only minimally efficient (about 1%) and hence after such an experiment majority of the cells do not have the foreign DNA. However the use of selectable marker genes which are an integral part of any cloning vector makes the selection of transformed cells quite easy. Generally, the selection methods are based on the expression or non-expression of certain traits such as antibiotic resistance, expression of an enzyme such as β-galactosidase or protein such as GFP (Green Fluorescent Protein) and dependence or independence of a nutritional requirement such as the amino acid leucine. For example if the host *E. coli* cells have taken up the plasmid pBR322 then these cells will grow in media containing the antibiotics ampicillin or tetracycline whereas normal *E. coli* cells will be killed by the antibiotics. Thus only transformed cells, however few, will be selected for growth and division.

The simplest method for selecting the transformants relies on the presence of antibiotic resistance genes on the plasmid or phage based vectors as already discussed. It is possible however, that the transformants have the vector without the foreign DNA. This is because the procedure for making a recombinant vector is efficient but not 100% foolproof! If on the other hand a vector has two antibiotic resistant genes, e.g. pBR322, and the insert is contained in the tetracycline resistant gene, then the ampicillin resistant gene will be normally expressed allowing the transformed cells to grow on an ampicillin containing medium but due to a phenomenon called insertional inactivation (insert in tetracycline gene) the cells will be tetracycline sensitive. How does a scientist select for a sensitive or negative trait following a transformation experiment? A procedure called **replica plating** is used. As schematically explained in **Fig. 7** following an experiment of transforming *E. coli* cells with recombinant pBR322 plasmid, the host cells are first plated on solid media (agarose containing) with the antibiotic ampicillin (assume that the insert has been ligated within the tetracycline resistance gene). Colonies from every single cell plate having the plasmid will develop overnight. The role of the recombinant plasmid is to help the cell to multiply in the presence of antibiotic, which it would otherwise not be able to do. In order to select those colonies alone which are tetracycline sensitive and therefore, are relevant to the experiment as they have the insert, a procedure called replica plating is used. A petri plate containing solid media with antibiotic tetracycline is kept carefully under asceptic conditions(Laminar flow hood) to which a circular piece of velvet or velvet paper is aligned and pressed onto the colony containing ampicillin plate (master plate). With the same alignment it is pressed onto the tetracycline plate. Overnight only colonies not containing the insert will grow while due to insertional inactivation no colonies will grow which have the insert. The colonies which have the insert can easily be scored off by comparing the two plates.

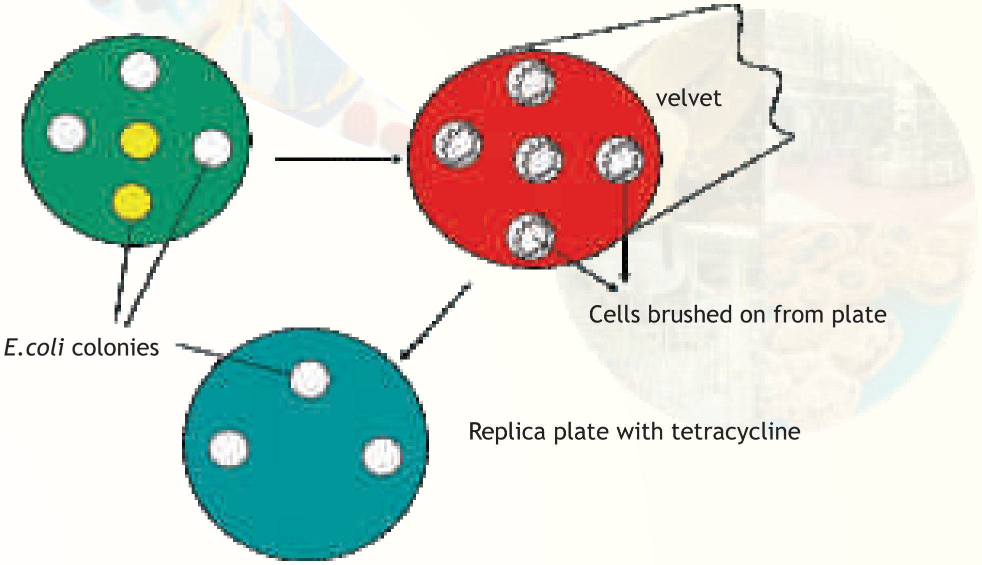


Fig. 7. Replica plating (note only colonies marked yellow have insert).

Another powerful method of screening for the presence of recombinant plasmids is referred to as blue- white selection. This method is based upon the insertional inactivation of the *lac Z* gene present on the vector (*e.g.* pUC 19). This gene expresses the enzyme β-galactosidase whose activity can cleave a colourless substrate called X-Gal into a blue coloured product. If the *lac Z* gene is inactivated due to the presence of the insert then the enzyme is not expressed. Hence if after a transformation experiment the *E. coli* host cells are plated on an ampicillin and X-Gal containing solid media plate then colonies which appear blue are those which have transformed cells (antibiotic resistant) but do not have the insert (express active enzyme). Colonies which appear white are both ampicillin resistant and have the insert recombinant DNA and thus are the cells to be used for future experiments.

The above described methods are used for selection of *E. coli* recombinants. There are several methods used for other host cell recombinants but the principles remain the same. Furthermore techniques for actually detecting recombinant proteins from colonies also have to be used where relevant. Where amplification of the insert DNA is the primary objective, plasmids are isolated from the host cells after growing the latter in large amounts and using the same restriction enzyme the insert is cut from the plasmid and recovered after electrophoresis.

5.1.6. Polymerase Chain Reaction (PCR)

The polymerase chain reaction or PCR as it is commonly known, was invented by Kerry Mullis in 1985. It results in the selective amplification of a specific region of a DNA molecule and so can also be used to generate a DNA fragment for cloning. The basic principle underlying this technique is that when a double-stranded DNA molecule is heated to a high temperature, the two DNA strands separate giving rise to single stranded molecules which can be made to hybridise with small oligonucleotide primers (single-stranded) by bringing down the temperature. If to this an enzyme called DNA polymerase and nucleotide triphosphates are added, much like what happens during replication, i.e primer extension occurs. This procedure is repeated several times, see **Fig. 8** which ultimately results in amplification of the DNA stretch between the two primers (one on each strand of the DNA). The basic requirements of a PCR reaction are:

1. DNA template to be amplified.
2. Primers which are oligonucleotides, usually 10-18 nucleotides long that hybridise to the target DNA region one to each strand of the DNA. Two primers of such a sequence are required so that they can hybridise as indicated in the **Fig. 8**.
3. DNA polymerase which is stable at temperatures above 80°C . Taq polymerase which has been isolated from a thermostable bacterial species is used.
4. Deoxynucleotide triphosphates and buffer.

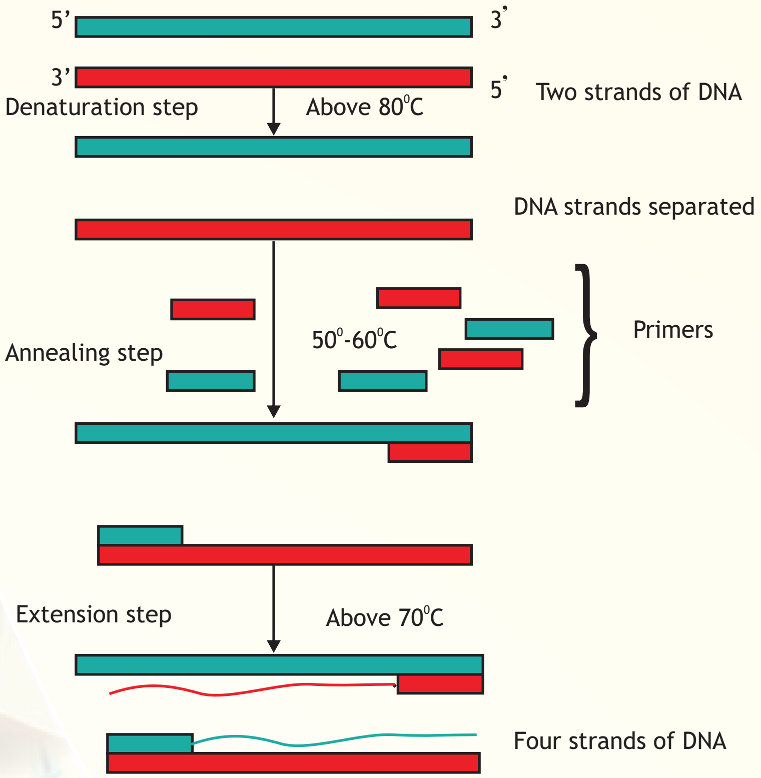


Fig. 8. PCR technique

A single PCR amplification cycle involves three basic steps of denaturation, annealing and extension **(Fig. 8)**. In the denaturing step, the target DNA is heated to a high temperature, above 80°C which results in DNA strand separation. Each single strand then anneals with a primer at a lower temperature between 50-60°C in such a way that extension can occur from it in a 5' - 3' direction (a requirement of the DNA polymerase). The final step is extension, wherein the enzyme Taq polymerase extends each primer using dNTPs and the DNA strand as template. The temperature for extension is around 70°C. The procedure is repeated and each set of steps is considered as one cycle (i.e denaturation, annealing and extension). At the end of one cycle two DNA molecules have become four and this geometric progression occurs with each cycle. Hence it can be computed that at the end of n cycles the number of DNA molecules is 2^n which is enormously amplified target DNA.

The invention of the PCR technique has revolutionised every aspect of modern biology. To detect pathogens, microbiologists in the past used techniques based on culturing and detecting antibodies against enzymes or proteins specific to the pathogen. Apart from taking time many of these procedures were not specific. PCR based diagnosis is faster, safer and more specific because it does not use live pathogens; instead DNA from the infected tissue is isolated and the PCR technique is carried out using primers having specific complementary sequences to the pathogen DNA. PCR is also a valuable tool in forensic science as large amounts of DNA can be amplified from the small amounts present at the crime site, for DNA fingerprinting analysis. In recent years, PCR has also found use in detecting specific microorganisms from environmental samples of soil, sediments and water. It is interesting that archaeologists are using combinations of PCR and fingerprinting analysis to relate and establish ancient Egyptian dynasties from samples obtained from mummies.

5.1.7. Hybridisation Techniques

Once a specific DNA sequence has been isolated by cloning, it can be used as a probe to detect the presence and amount of complementary DNA sequences present in isolated DNA, for example from different species. A question can be asked such as, Is this particular gene or DNA fragment present in a mammalian cell also present in an insect cell?

DNA probes are relatively small single stranded sequences of DNA that recognise and bind to complementary sequences. Recall that two strands of DNA are held together by base complementarity *i.e.,* base A on one DNA strand hydrogen bonds with base T on the complementary strand and likewise base C hydrogen bonds with base G. A DNA probe which is single stranded will bind to a complementary sequence with the same base pairing rules (hybridisation) and if the probe can be tagged with a flourescent or radioactive label the complementary sequence can be located either in a cell nucleus or on a gel chromatogram. This principle is the basis of all hybridisation techniques, see **Fig. 9.**

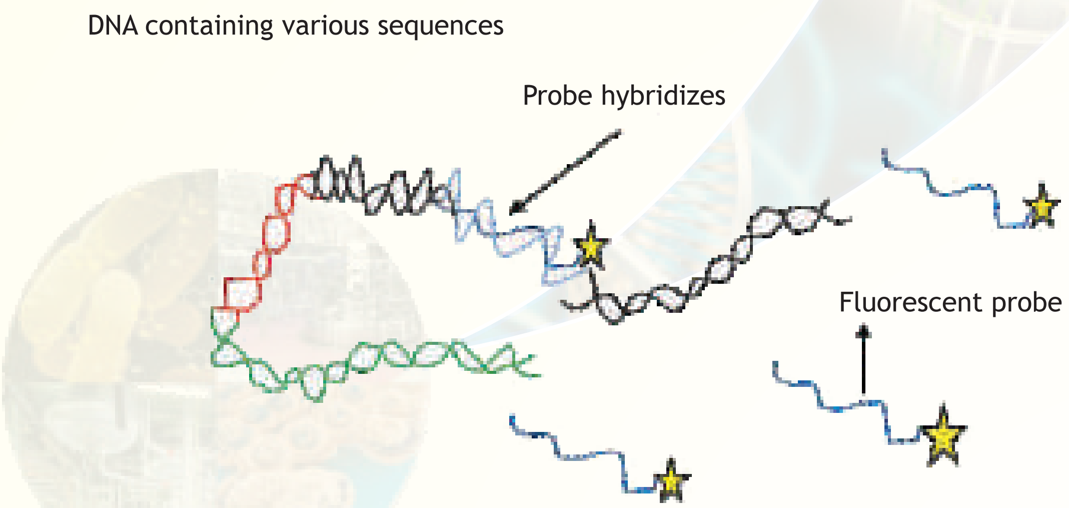


Fig. 9. Principle of hybridisation technique.

If the probe is tagged with a fluorescent label, under UV light its location can be easily seen as it will fluoresce. On the other hand if the probe is tagged with a radioactive label a technique called autoradiography is used wherein the gel is placed on a photographic film and the probe location is indicated by white spots on the developed film.

Southern Hybridisation Technique

This technique of identifying and locating specific sequences in DNA gels using probes was invented in 1975 by Edward Southern and is named Southern Hybridisation technique in his honor. This is an essential technique in all rDNA experiments and its aim is to identify a specific DNA sequence in a heterogenous population of DNA molecules. As discussed in the previous section the principle of the technique is based on the ability of a probe to seek out and bind to its complementary sequence.

The procedure involves isolation and digestion of total genomic DNA with one or more restriction enzymes. The DNA fragments thus generated are separated in agarose gels using the technique of electrophoresis, about which you have learnt from Class XI Biotechnology textbook. Following separation of the DNA fragments due to size differences, they are transferred from the gel to a nylon or nitrocellulose membrane in a technique called blotting, see **Fig. 10**. In the blotting procedure the DNA fragments are forced from the gel onto the membrane by capillary action. The membranes are baked briefly to fix the DNA fragments to the membrane so that they do not diffuse during the next step of hybridisation. Note that the membrane will have the same pattern of DNA bands as the original agarose gel. The membrane is then treated with the single stranded labelled probe for an appropriate period after which the membrane is washed and either photographed under UV light (if probe label is fluorescent) or overlaid with a photographic film (if probe is radioactive).   The location of the probe is determined leading to the identification of a gene or specific DNA fragment obtained from that given genomic DNA. A similar principle is used for determining RNA locations and this technique is called Northern hybridisation.

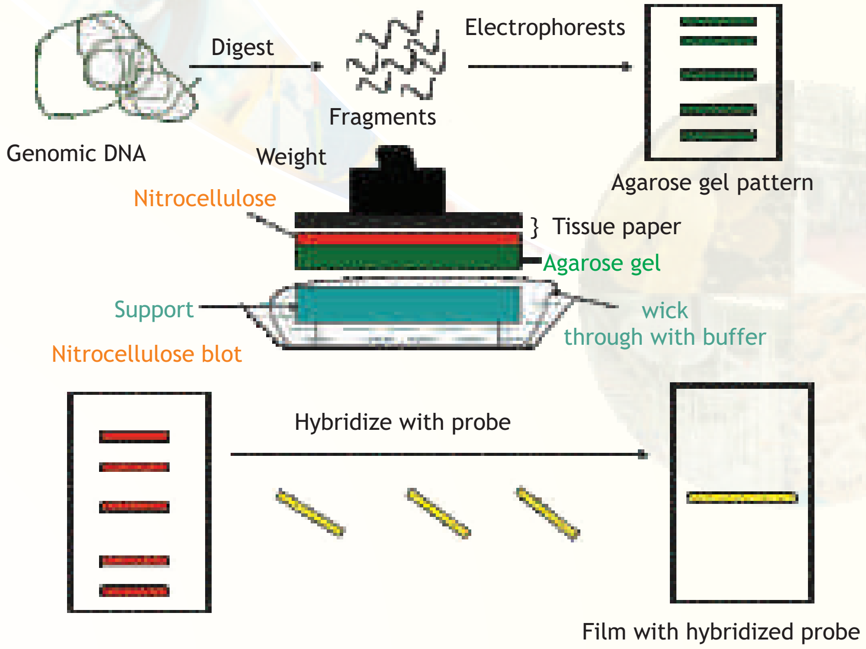


Fig. 10. Southern hybridisation technique.

5.1.9. DNA Library

In order to obtain a DNA fragment carrying a useful target gene it is essential to make a DNA library containing all possible DNA fragments from the genome of a given species or cell type. The target gene can then be easily identified from the library and then used for cloning. Remember per cell only two copies (diploid) of a target gene are present and therefore too little is there for detection and use. A DNA library not only contains all possible fragments of DNA from a given cell or organism but also large amounts of the same as a resource. Two types of DNA libraries can be constructed - a genomic DNA library and a cDNA library. A genomic DNA library has all possible DNA sequences in large amounts from the given cell type. A cDNA library on the other hand has only expressed gene sequences such as protein encoding genes. It is obvious that a genomic DNA library is larger than a cDNA library. Whether the fragments are genomic or cDNA (protein encoding) these are inserted into vectors such as plasmids and then introduced into *E. coli* hosts. The cells are diluted and plated on agarose plates so that distinct colonies are formed starting from single cells. Each colony would therefore have an amplified amount of a given fragment. However unlike an *E.coli* regular library which has a classification system, a DNA library has to be screened for finding a fragment or gene of interest which is a tedious and time taking process. A schematic diagram of the steps in constructing a DNA library is indicated in **Fig. 11**.

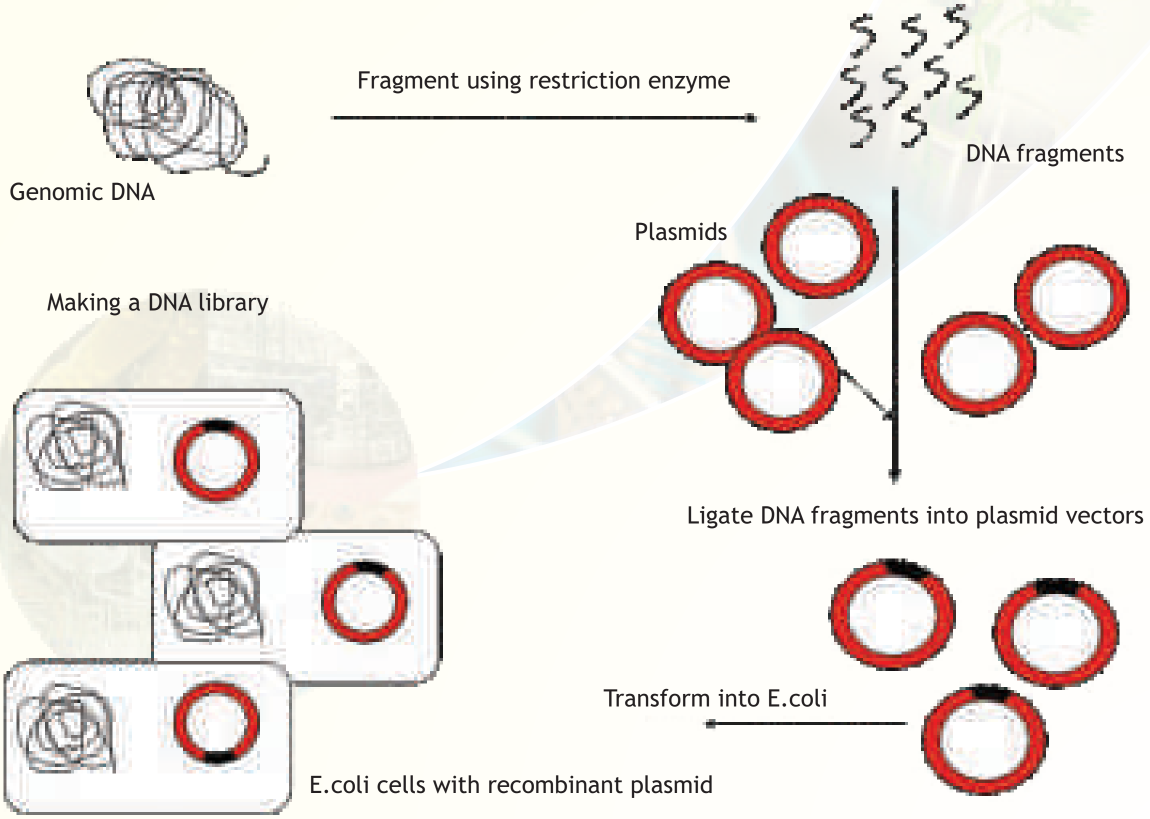

Fragment using restriction enzyme
Genomic DNA
DNA fragments
Plasmids
Making a DNA library
Ligate DNA fragments into plasmid vectors
Transform into E.coli
E.coli cells with recombinant plasmid

Fig. 11. Making a DNA library.

As shown in **Fig. 11**, to prepare a genomic library, the total genomic DNA is isolated from a tissue or organism and then fragmented using a restriction enzyme such as *Eco*RI. An appropriate vector such as pBR322 (plasmid based) is also digested with the same enzyme and is then treated with the enzyme alkaline phosphatase which as indicated elsewhere removes the 5'phosphate group to prevent the plasmid from self ligation. The DNA fragments and cut vector are mixed and then treated with the enzyme DNA ligase. Each vector molecule will contain a different fragment of DNA and these are introduced into *E. coli* host cells by a technique called transformation. More details about this technique will be discussed in the following section.

Although a genomic library represents the entire genome of an organism, it may not be useful in case of eukaryotic organisms. Genomic DNA of eucaryotes contains non-coding DNA like introns, control regions and repetitive sequences unrecognised by prokaryotic systems such as in *E. coli*. Therefore another library called cDNA library is preferred especially if the aim of the cloning experiment is to express eukaryotic proteins in an *E. coli* host. This library has two major advantages over a genomic library. Firstly, it represents only those genes that are being expressed by a particular cell or under specific conditions. Eucaryotes, which include multicellular organisms such as animals and plants have differentiated cells which means that different cells e.g. liver and brain cells express some common proteins and also other proteins which are not common. Secondly, since the source material in constructing such libraries is mRNA, these molecules lack introns and hence would represent only the coding sequences of the genome. However mRNA molecules are highly unstable as they are easily degraded by RNAses and hence these molecules are faithfully copied into the more stable DNA (now called cDNA) before cloning. The construction of a cDNA library begins with the isolation of mRNA from a given cell type or tissue which are copied into cDNA using a special enzyme called **reverse transcriptase.** The procedure results in double-stranded cDNA which can be incorporated into vectors such as pBR322.

5.1.9. DNA Sequencing

In this era of genomics wherein whole genomes of species are being sequenced and compared to get a vision into the fundamental nature of DNA, the blueprint of life, the ease with which DNA is sequenced has played a major role. In the seventies two major methods were developed to sequence DNA:

1. The dideoxynucleotide chain termination method invented by Fred Sanger (the same scientist who invented a protein sequencing) and Andrew Coulson

2. Chemical degradation method invented by Walter Gilbert and Allan Maxam (also called the Maxam and Gilbert method).

   Of the two methods the first method is more popularily used and hence this technique will be discussed.

Dideoxynucleotide Chain Termination method

In your class XI textbook you would have read about replication of DNA and also about the various enzymes and substrates required. DNA polymerases, the major enzymes required in replication, have certain properties:

- A single stranded DNA template is required for them to act upon.
- A new strand cannot be initiated; only primers can be extended using the single strand DNA template as a guide.
- Extension or DNA synthesis occurs in a 5'→ 3' direction which requires that a new nucleotide is added to the 3' hydroxyl group of the chain.
- Deoxynucleotide 5' triphosphates are the normal substrates.
- However if the 3' hydroxyl group of a deoxynucleotide triphosphate is absent as in the 2',3' dideoxynucleotide triphosphate **(Fig. 12)** it would result in, that nucleotide being incorporated into the growing chain but subsequently the chain cannot further be extended (no 3' hydroxyl) and this is the main principle of the Sanger Dideoxynucleotide Chain Termination method.

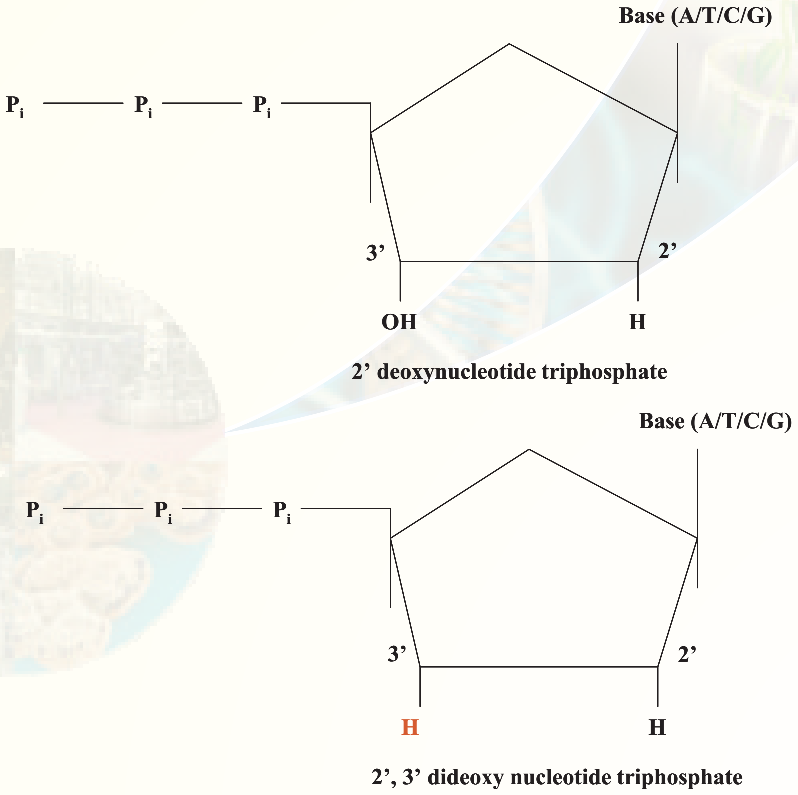
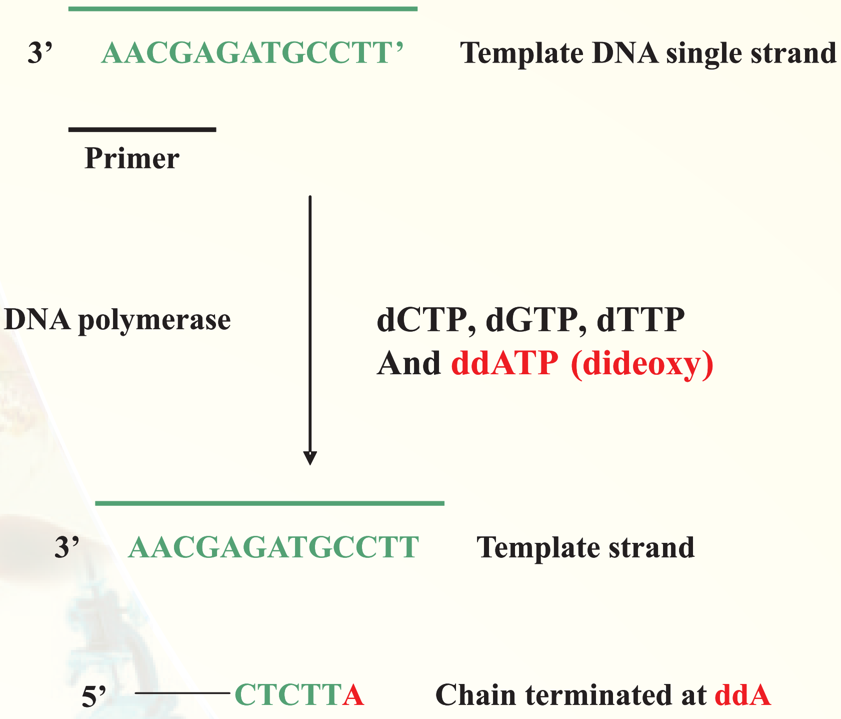


Fig. 12. Deoxy and dideoxy nucleotides.

During a primer extension process, whenever a dideoxynucleotide is incorporated, the DNA polymerase enzyme cannot further carry out its reaction and the new DNA strand is thereby terminated **(Fig. 13)**



3’ AACGAGATGCCTT’ Template DNA single strand

Primer

DNA polymerase

dCTP, dGTP, dTTP
And ddATP (dideoxy)

3’ AACGAGATGCCTT Template strand

5’ CTCTTA Chain terminated at ddA

Fig. 13. Principle of dideoxynucleotide chain termination method.

Typically the sequencing technique is carried out in four test tubes. Each test tube carries the following common reagents, single stranded DNA template, deoxynucleotide triphosphates, primer and DNA polymerase. In addition to it a small amount of the four dideoxynucleotide triphosphates i.e ddATP, ddCTP, ddGTP and ddTTP is added separately into the four test tubes. The purpose of adding only a small amount of the dideoxy derivative is that only one dideoxy derivative is incorporated into each extending chain causing termination. Hence in a test tube containing ddATP all chains will end at ddA but at different positions of T present in the template **(Fig. 14)**. The same is applicable to the other test tubes containing ddCTP, ddGTP and ddTTP. It is to be noted that the dideoxynucleotide is added wherever the complementary base is present on the template.

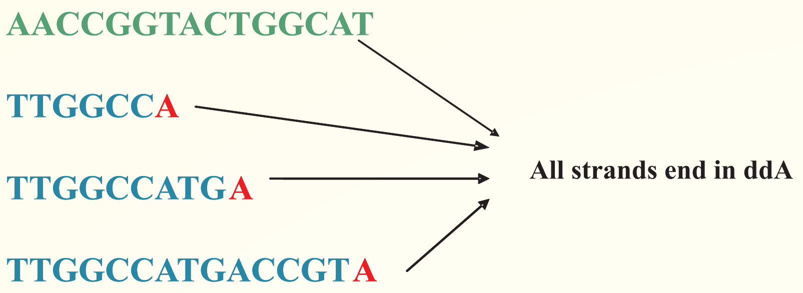


Fig. 14. Principle of chain termination at a particular dideoxynucleotide.

The various strands prematurely terminated at the particular ddNTP in a given tube are subjected to electrophoresis in special gels wherein bands can be resolved (separated) even if they differ by one nucleotide. The strands migrate in the gel with the shorter fragments moving faster towards the anode. The primers used in each tube can be made radioactive and hence the position of the separated strands in the gel chromatogram can be easily visualised using autoradiography. **Fig. 15** is a schematic representation of a typical gel and its read sequence. Colours for different dideoxy nucleotides are for clarity.

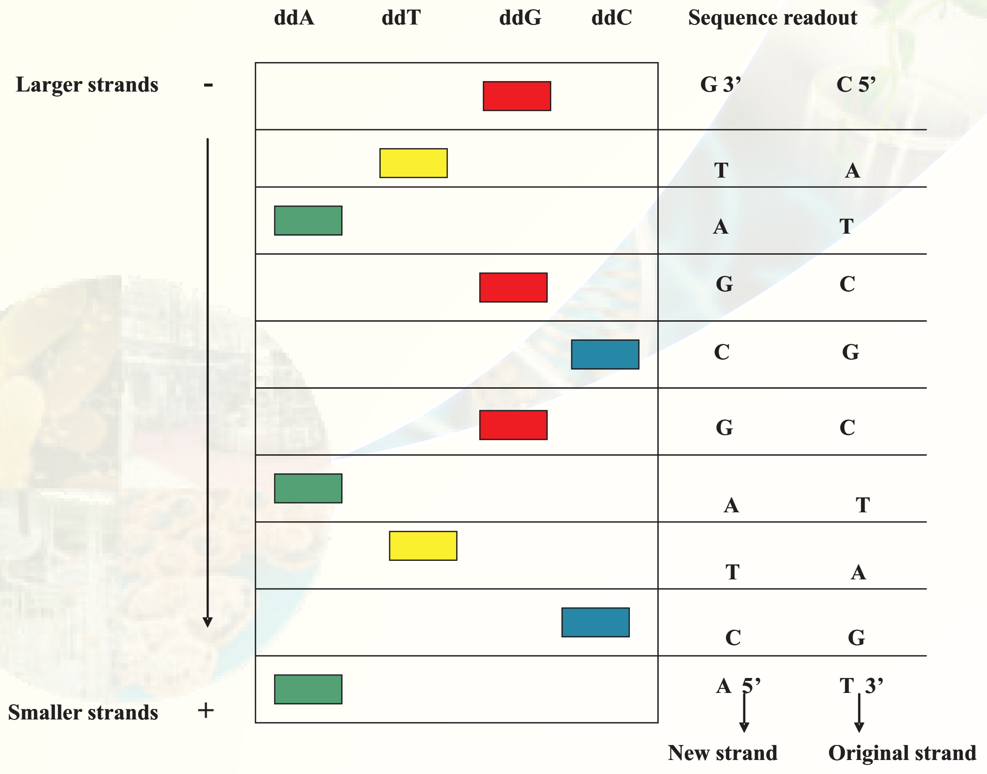


Fig. 15. Reading a sequencing gel pattern (lines drawn for clarity in reading).

Nowadays DNA sequencing technologies have become automated. To avoid using radioisotopes and their consequent danger, dideoxynucleotides are conjugated with florescent molecules which on excitation give a different colour each. Hence each band on the gel (read from anode to cathode) indicates the particular base as its terminal dideoxy nucleotide fluoresces with a given colour. This avoids the use of a four lane gel, a single lane gel electrophoresis is instead conducted and the gels are then laser scanned and the data fed into a computer. The computer is programmed to display the gel scan and the base readout as shown in **Fig. 16.**

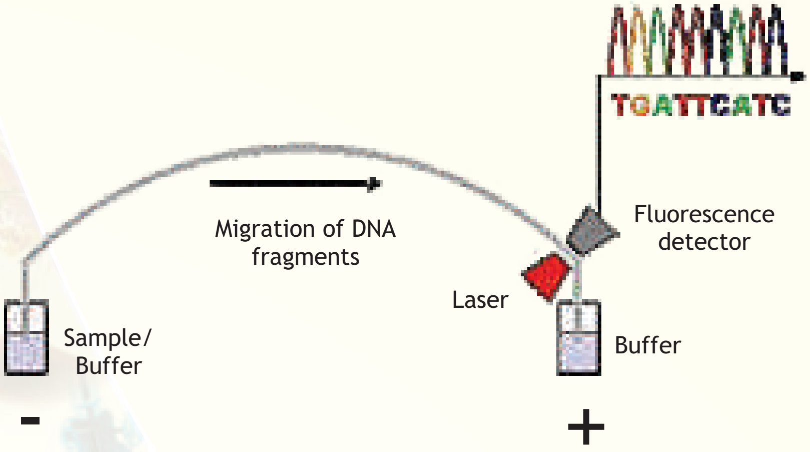


Fig. 16. Single Tube Sequencing experiment.

5.1.10. Site-directed Mutagenesis

Mutation is an alteration in any of the base of a DNA sequence sometime's leading to a defective protein or prematurely terminated non-functional protein. Mutations are spontaneous in nature although rare (*e.g* sickle cell haemoglobin). In the technique of site-directed mutagenesis a Biotechnologist is able to create mutation selectively, rather than that which occurs randomly in nature. Using this technique amino acids can be substituted in the expressed proteins making them more stable or functionally better. Furthermore the role of specific amino acids in proteins has led to a better understanding of protein structure and function.

The principle of site-directed mutagenesis as schematically explained in **Fig. 17** involves cloning the target gene into an M13 vector wherein it is presented as a single stranded part of the phage genome. A small oligonucleotide is added containing a complementary sequence to the gene but with one or more altered nucleotides. This allows the oligonucleotide to bind to a complementary portion in the target gene. This then acts like a primer *in vitro* to synthesise a double stranded replicative form. Note that the duplex RF form has one strand with the original target gene sequence, wild type and the other strand with the altered nucleotide(s). The duplex DNA molecule is then introduced into bacterial cells by transformation. Subsequent replication inside bacterial cells will produce either wild type or mutant gene containing plasmids. If appropriate expression signals are present altered protein can be expressed and studied. In the unit on Protein engineering a site-directed mutagenesis experiment has been described wherein a stable form of the proteolytic enzyme subtilisin has been generated which is used in laundry detergent.

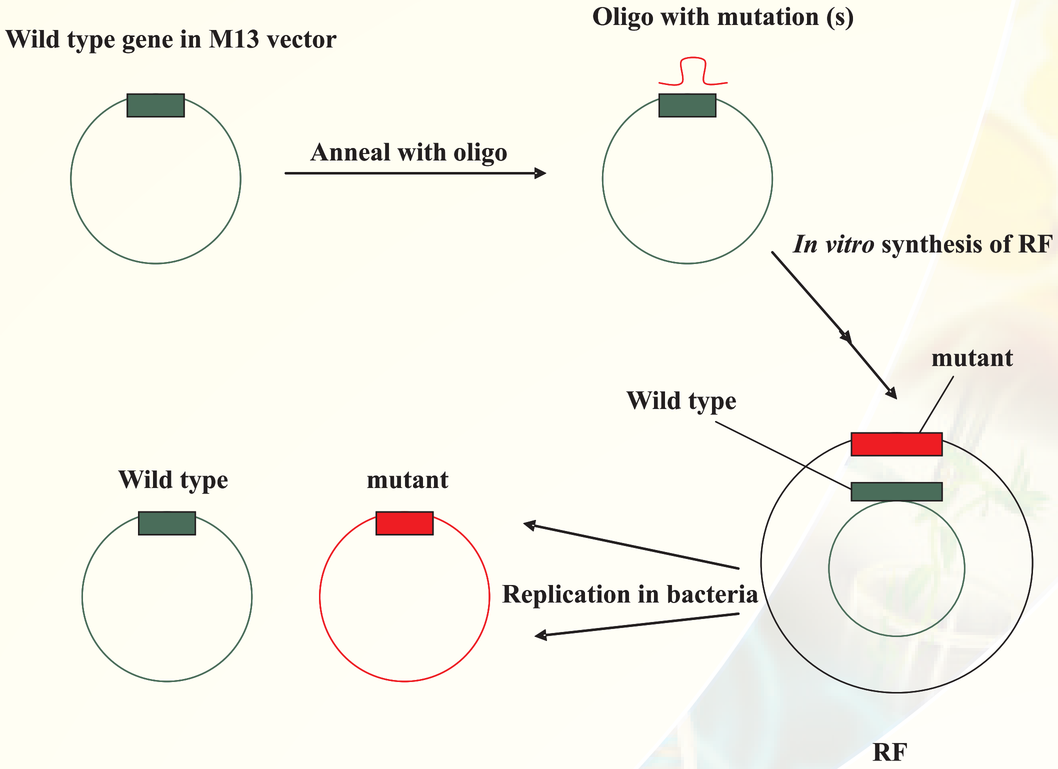


Fig. 17. Site-directed mutagenesis

Review Questions

1. Define the following:

Plasmid	Restriction site
Transformation	Mutation
Transfection	

2. What are restriction enzymes and why are they so important in rDNA technology?

3. Enlist the various steps involved in a rDNA experiment.

4. What are the essential features of a vector?

5. What does PCR stand for? Name the different steps in a PCR reaction.

6. Explain 'Insertional Inactivation'.

7. What are the disadvantages of using E. coli for production of eukaryotic proteins?

8. Distinguish between:

   i) Blunt ends vs sticky ends

   ii) YAC vs BAC

   iii) Genomic library vs cDNA library

   iv) Microinjection vs Electroporation

9. Write a short note on RFLP and indicate one of its important applications.

10. Why are ddNTPs used in sequencing? Briefly indicate the principle of DNA sequencing using these.

11. Indicate one application of site-directed mutagenesis.

References

1. Molecular Biotechnology by S.B. Primrose, Panima Publishing Corporation, New Delhi (1999), 2nd Edition.

2. Molecular Biotechnology by Glick and Pasterneck, Panima Publishing Corporation, New Delhi (1999), 2nd Edition.

3. An Introduction to Molecular Biotechnology Edited by Michael Wink, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany (2006).

4. Principles of Gene Manipulation and Genomics by S. B. Primrose and R. M. Twyman, Blackwell Publishing, Carlton, Victoria, Australia (2006), 7th Edition.

5. From Genes to Clones by Ernst-L. Winnacker, Panima Publishing Corporation, New Delhi (2003).

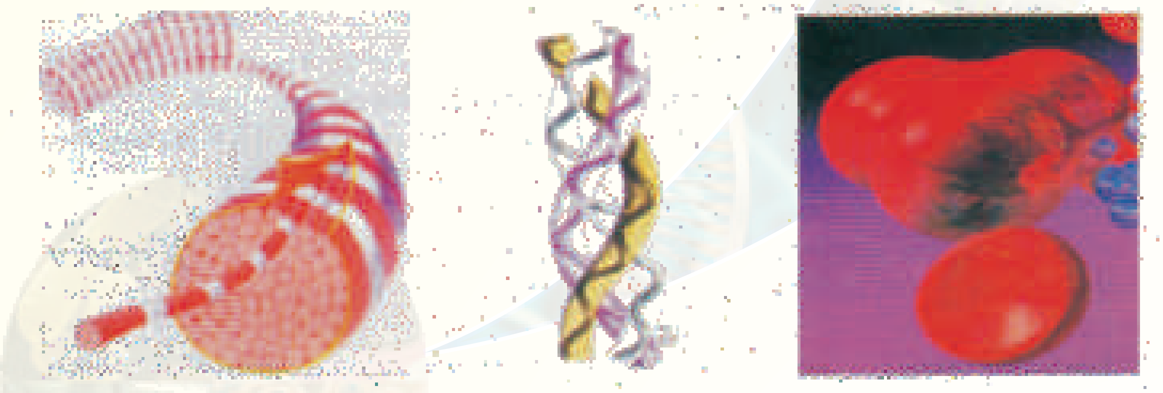
6. Gene Cloning and DNA Analysis: An Introduction by T. A. Brown, Blackwell Sciences Ltd., Oxford (2001), 4th Edition.

CHAPTER 2

PROTEIN STRUCTURE AND ENGINEERING

5.2.1. Introduction to the World of Proteins

The shape, structure and function of the human body is one of the Nature's marvels. The fertilisation of an egg by a sperm to the growth of a whole human body involves numerous steps of growth and differentiation. When we breathe, we feel a sense of oxygen flowing through our lungs and racing in our blood vessels, to be delivered to all our tissues. While we flex our muscles, we can feel them first tightening and then relaxing. The molecules involved are proteins, haemoglobin which transports oxygen, collagen which provides the strength to our bones and extracellular tissue and actin, myosin and several others which help in muscle contraction **(Fig. 1)**. It is noteworthy that among the biomolecules you have studied, proteins have the maximum diversity in function. The key to this enormous diversity is the unique structure of proteins. Although all proteins are made up of 20 different amino acids the sizes and sequence combinations and variations of each protein leads to millions of unique 3-D structures and thereby functions. Scientists have been striving to relate protein structure with function and hence the first step would be to determine 3-D structure of a protein.



Myosin and actin

Collagen

RBC with Haemoglobin

Fig. 1. Proteins having multiple roles

Even more amazing is the structure and data processing abilities of the human brain. We tend to marvel at the incredible speed and processing functions of the super computers little realising the creativity of the human being who invented them. The speed and correlation of sensory stimulation is unique to the human brain which can grasp the diversity of sensory inputs and convert them to learning and memory for later application. What are these proteins which enable these functions and why are some brain related diseases like Alzeimers occurring, in which certain proteins show abnormal structure and behaviour?

A number of human diseases are due to the deficiency or abnormal structure of proteins. The lack of a particular subunit, alpha or beta of the oxygen carrying protein haemoglobin results in Thalassaemia, a devastating disease in which an infant cannot grow without repeated transfusions. If the beta chain is present but with a substituent in one of the amino acid residues another debilitating condition called Sickle Cell Anaemia results which is endemic to certain parts of Africa. The absence of an enzyme- Adenosine deaminase results in the birth of a severely immunocompromised baby who cannot last infancy (SCID). More recently it has been discovered that certain "rogue proteins" whose structure has been altered can result in diseases such as the Mad cow disease wherein the disease itself appears to be propagated by infectious proteins called **"prions"**. Clearly proteins need to be understood in detailed terms.

The completion of the Human Genome Sequence has revealed about 35,000 genes. However the actual number of proteins encoded by these genes may be many more due to posttranscriptional modifications. Different cells have specialised proteins for their unique functions in addition to the housekeeping proteins required for metabolism and generation of ATP. Sometimes these proteins are secreted to the outside like the proteolytic enzymes from the pancreas or hormones from ductless glands like the pituitary. We are yet to identify all the proteins required for a body to function and this presents a challenge to the future biotechnologists. One of the outcomes is the merging field of protein structure and function- proteomics. This chapter will enable you to understand various features of the area of proteomics- 3-D structure, functions and applications of protein products, some generated by biotechnological processes.

5.2.2. 3-D Shape of Proteins

The morphology, function and activity of a cell are all dependant on the proteins expressed. Proteins perform a variety of roles. The three dimensional properties of proteins have an important bearing on their function. The first step in determining the structure of a protein is to isolate it in a pure form from its cellular location (note bacterial, plant or animal cell). The purified protein is then crystallised so that using a technique called X-ray crystallography its three dimensional structure can be deduced. Nowadays another powerful technique called Nuclear Magnetic Resonance (NMR) has been developed which can deduce protein structures in solution and hence crystallisation is not required. However the protein in either technique has to be purified and some general procedures used to purify proteins will be discussed in subsequent sections of this chapter. In general when we refer to the structure of a protein this involves two aspects- the chemical structure which is the amino acid sequence of the polypeptide and its folding in space which is referred to its 3-D structure.

One of the major breakthroughs in protein sequence determination was achieved in the middle of the last century by Dr. Frederick Sanger who developed the first sequencing reagent FDNB (fluoro dinitro benzene) and a general strategy for sequencing. By using these methods he was able to sequence the important hormone insulin which is required by diabetics and more importantly he demonstrated for the first time that proteins were linear polymers of amino acids. For this work he was awarded the Nobel Prize and it will be interesting for you to know that several years later he was awarded a second Nobel Prize for developing a sequencing technique for DNA which has been described in the Recombinant DNA technology chapter previously. Another protein chemist, Pehr Edman in 1950 developed another sequencing reagent and procedure which is used in modern day sequenators as the procedure has been automated. These methods have been discussed in the XIth class' textbook of Biotechnology in Unit II, Chapter 2. Notably using the sequence of insulin established by Sanger a biotechnology company called Eli Lilli was able to develop recombinant human insulin which is the major source for insulin administration to diabetics worldwide.

With the availability of pure proteins, scientists like Linus Pauling, G.N.Ramachandran, Max Perutz and John Kendrew to name a few started developing techniques to study the 3-D shapes of proteins using high resolution X-rays. They laid the foundation for deducing protein structure by enunciating the basic rules which govern protein folding and the forces which cause the folding and stabilise them. Hence from these studies the concepts of planarity of the peptide bond, secondary structures such as alpha helix and beta pleats were developed. These concepts were introduced in the Class XI Biotechnology textbook.

Let us reiterate some important points regarding protein structure from the Class XI Biotechnology textbook. Protein structure has been divided into four hierarchial levels to understand their organisation:

The linear order or sequence of covalently linked amino acid sequence is defined as **primary structure**. Depending on the nature and arrangement of the amino acids present different parts of the polypeptide chain form **secondary structures** like alpha helices and beta pleats. The **tertiary structure** organisation of these secondary structural elements occurs when these get compacted with each other to form compact spherical or globular units which are also thermodynamically stable conformations of these molecules in aqueous solutions (note cytoplasm is mainly water). In compaction several non-covalent interactions occur between the amino acid side chains. The **quarternary structure** is the association of two or more independent proteins/polypeptides via non-covalent forces to give a multimeric protein **(Fig. 2)**. The individual peptide units of this protein are referred to as subunits and they may be identical or different from one another.

The dominant forces which cause linear protein chains to undergo folding in space lies to a large extent in the chemistry of the amino acid residues they contain. Amino acids are broadly divided into three main groups- **polar or hydrophilic** (eg. serine, glutamine), **charged** (eg. aspartate, arginine) and **hydrophobic** (eg. tryptophan, valine). Hence based on these features amino acid side chains can interact in space by a variety of non-covalent forces which is the basis of forming and stabilising protein structures in space. Let us examine some of the major non-covalent forces found in proteins.

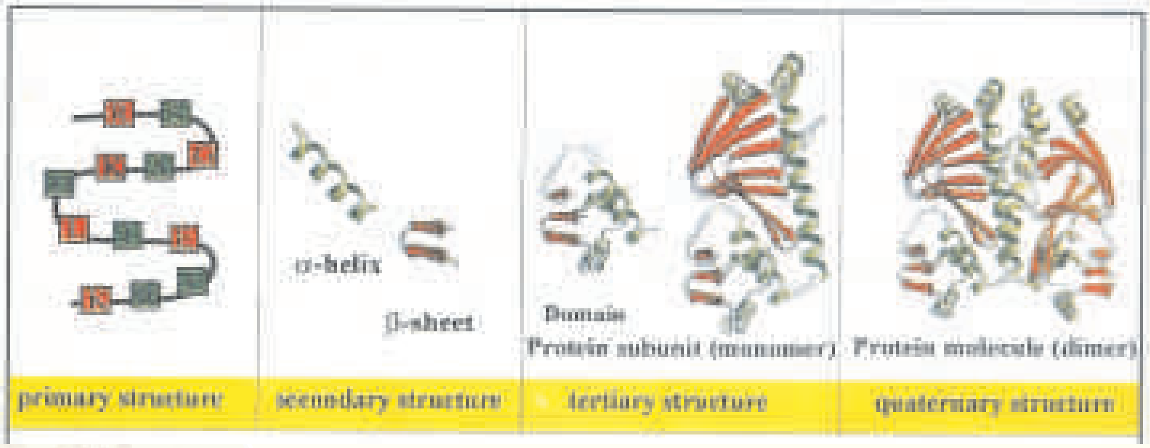


Fig. 2. Hierarchical organization in protein structure

Non-covalent bonds

The non-covalent interactions involved in organising the structure of protein molecules can be broadly divided into four categories:

- Ionic bonds
- Hydrogen bonds
- Van der Waals forces
- Hydrophobic interactions

Ionic bonds

These involve interactions between the oppositely charged groups of a molecule. For example the positively charged amino acid side chains of lysine and arginine can form salt bridges with the negatively charged side chains of aspartate and glutamate. These ionic interactions are also known as **salt bridges** because these are dominant bonds found in salts like sodium chloride wherein the positively charged sodium ion interacts with the negatively charged chloride ion. However, although ionic bonds have similar strengths to covalent bonds in vacuo, the bond strength of ionic bonds is vastly reduced in water due to the insulating qualities (dielectric strength) of water. Ionic bonds are highly sensitive to pH and salt concentration.

Hydrogen bonds

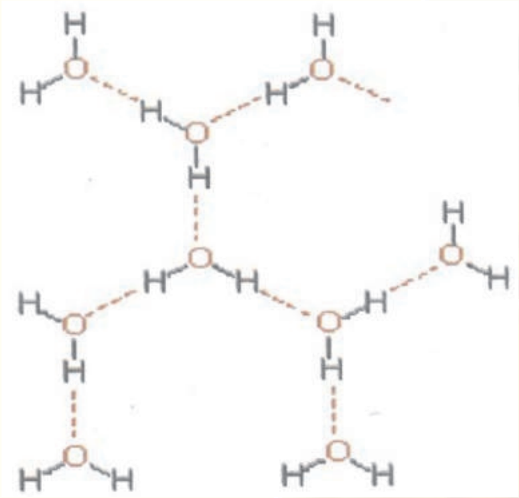


Fig. 3. Hydrogen bonding network in water

Hydrogen bonds are formed by "sharing" of a hydrogen atom between two electronegative atoms such as Nitrogen and Oxygen. In this case strongly polarised bonds between hydrogen and a small, very electronegative atom (N,O or F) allow a strong dipole-dipole bond to be formed with another small very electronegative element (N, O or F, **Fig. 3**). Importantly, the very small sizes of these elements also allow them to approach each other so closely that a partial covalent bond is also formed (*e.g.*O-H---N). It is to be noted that the partial covalent character means that these bonds (H-bonds) are directional and strongest when the nuclei of all three involved atoms are in a linear arrangement.

Van der Waals forces

These forces are weak attractions (or repulsions) which occur between atoms at close range. The Van der Waals types of forces are essentially contact forces, proportional to the surface areas in contact. These forces are of little significance at a distance due to the rapid 1/r6 (r is the inter-atomic distance) fall off. Even though weak, these bonds can be important in macromolecules because the large surface areas involved can result in reasonably large total forces.

Hydrophobic interactions

Hydrophobic interactions can be best explained by taking an example of oil in water. The oil tends to separate out fairly quickly, not because the oil molecules "want to get together", but because the water forces them out. The hydrophobic interaction is thus a manifestation of hydrogen bonding network in water. In water, each molecule is potentially bonded to four other molecules through H-bonds **(Fig. 3)**.

If a non-polar molecule, which cannot participate in hydrogen bonding, or in electrostatic interactions with water molecules, is added into water , a number of hydrogen bonds will be broken and not replaced. Since hydrogen bonds are favourable interactions, there will be an energy cost to putting non-polar molecules into water. Water therefore forces these molecules out of solution to minimise the surface of contact and thus the number of hydrogen bonds which are broken. Such forces known as hydrophobic forces are among the most important in driving proteins to fold into compact structures (globular) in water. Also, these forces are responsible to make different proteins assemble together to form structures found in muscles, membranes and other organs. In proteins therefore, hydrophobic regions are preferentially located away from the surface of the molecule and form the interior core of the protein.

5.2.3. Structure-Function relationship in Proteins

As you have now learnt about how various forces drive proteins to assume characteristic shapes, it is worthwhile to consider why shape is paramount to the function of a protein. We will look at two proteins- an enzyme, chymotrypsin and the oxygen carrying protein, haemoglobin, to emphasise the importance of protein structure in its function.

Chymotrypsin, a proteolytic enzyme

As injested food makes way into the duodenum from the stomach, the proteins encounter a fierce proteolytic duo- trypsin and chymotrypsin which precisely cut the linear chains into short peptides which later on are acted upon by peptidases to release amino acids. Chymotrypsin, which hydrolyses peptide bonds following bulky aromatic amino acid residues in polypeptides is actually synthesised in the pancreas and through the pancreatic duct released into the duodenum. Have you wondered why this enzyme being a powerful proteolytic enzyme does not end up cutting cellular proteins within the pancreas itself? Nature has ensured that chymotrypsin and other proteolytic enzymes are synthesised as inactive harmless precursors known as zymogens which are then activated when required only in the duodenum, their site of activity, a process called *in-situ* activation. This activation in molecular terms results in an alteration in its shape so that it may now be able to interact with its substrate. The inactive precursor enzyme is termed chymotrypsinogen and the fully active enzyme is called chymotrypsin. The enzyme chymotrypsin is made up of a linear chain of 245 amino acids interrupted into three peptides-A,B,C. The protein folds into a globular structure. In the 3-D structure of the enzyme three important amino acid residues, his57, asp102 and ser195 come close together in space **(Fig. 4)** which allows a "charge relay system" to operate as indicated in **Fig 5**. The negatively charged asp102 is able to hydrogen bond with the adjacent his57 partially borrowing the hydrogen ion from the latter. The his57 makes good its partial hydrogen ion loss to aspartate by attracting a hydrogen ion from the adjacent ser195 through the his57 residue much like a relay race where the baton is passed from one member to another, the difference here being that the baton is a charge.

Normally the hydroxyl group of a serine residue is not acidic (pKa 12) and this is true for all other serine residues of chymotrypsin; only ser195 becomes acidic due to the unique constellation of the three amino acid residues because the protein has folded uniquely in space. You may be curious about the importance about an acidic serine residue. The negatively charged oxygen anion is able to make a nucleophilic attack on the carbonyl carbon of the peptide bond of its substrate, loosening it so that a water molecule can hydrolyse the bond **(Fig. 5)**. The specific site of chymotrypsin (recall that the enzyme is specific to aromatic residues) is a large space created within the enzyme active site and lined by hydrophobic residues which therefore only allow bulky aromatic, hydrophobic amino acids to bind. This binding brings the susceptible peptide bond close to the attacking ser195 residue. In chymotrypsinogen, the substrate binding site is blocked and hence the enzyme is inactive. *In-situ* activation of trypsin involves a proteolytic cut in chymotrypsinogen which results in a conformational change, exposing the substrate binding pocket.

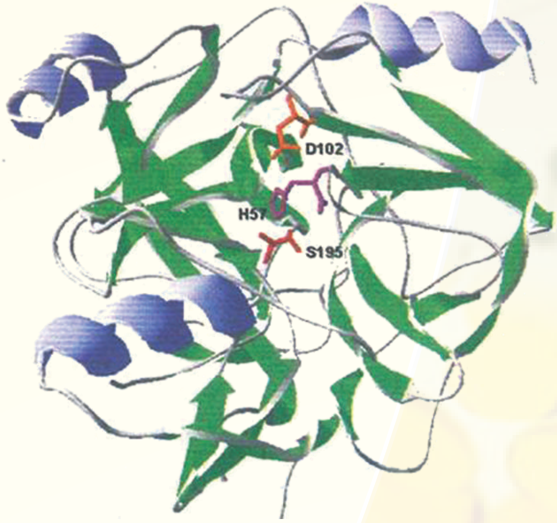


Fig. 4. Three dimensional structure of chymotrypsin

The interesting thing is that when nature has found a useful folding pattern which can cause hydrolysis of protein substrates, it repeats this in a variety of other enzymes. Trypsin, subtilisin (a proteolytic enzyme found in *B. subtilis*, a bacterium), thrombin (a proteolytic blood clotting factor) and the brain enzyme, acetyl choline esterase all have a reactive serine residue which is central to the catalytic mechanism.

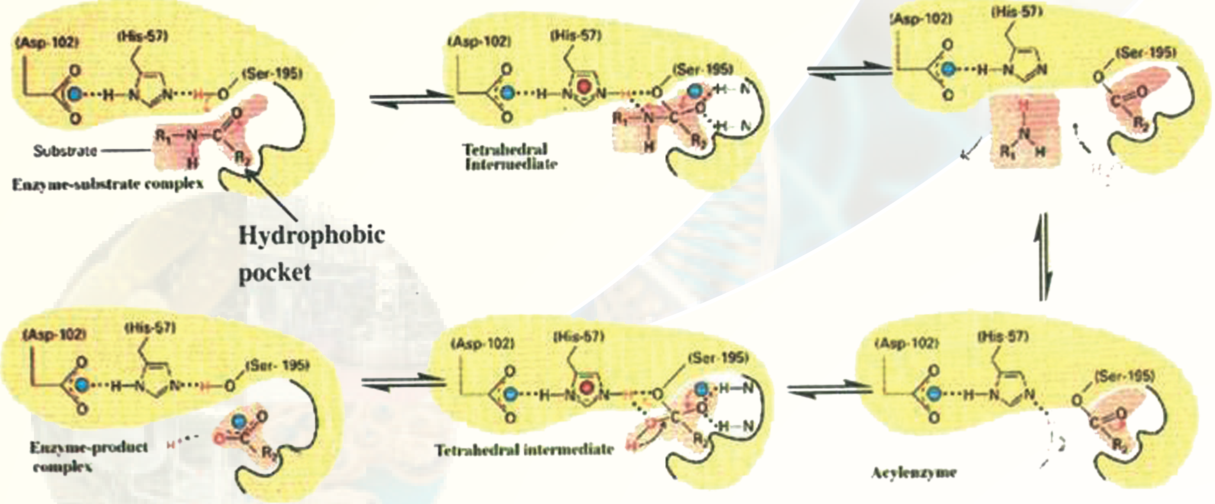


Fig. 5. Charge relay transfer in chymotrypsin. R_2 = aromatic amino acid; R_1 = any other amino acid.

Certain organophosphate compounds can selectively react with an acidic serine residue thereby knocking off enzyme activity. Nerve gas which was unfortunately used in the first world war had volatile serine alkylating compounds which inactivates the brain enzyme acetyl choline esterase leading to death. Nowadays derivatives of organophosphates such as malathion and parathion which are not toxic to humans are used as mosquito repellants (Mortein, Good Knight) by effecting nerve transmission in insects.

Molecular Disease- Sickle cell anaemia

Sickle cell anaemia is a disease prevalent in parts of Africa and India where malaria is also endemic. The red cells of the patient have a pronounced morphological change and resemble the shape of a farmer's sickle and thus the name of the disease. Because these unusually shaped red cells have impaired oxygen carrying capacity and further get stuck in the small capillaries they lead to the anaemic conditions observed in patients. Interestingly such sickled RBCs resist malarial infection and hence offer some selection unfortunately for malaria to be co-prevalent with sickle cell anaemia. One of the first attempts to study the molecular basis of sickle cell anaemia was to compare the electrophoretic mobility of normal (Hb) and sickle cell haemoglobin (scHb). On finding that Hb moved faster than scHb, Linus Pauling predicted that the latter differed in a charged amino acid. This was confirmed by V. M. Ingram in 1957 who pioneered a useful technique called protein finger printing in the famous Laboratory of Molecular Biology (LMB) at Cambridge, UK. LMB has been the Mecca for protein sequencing, DNA sequencing, X-ray crystallography, deduction of the Double helix structure of DNA, Hybridoma technology and Nematode developmental studies. Established in 1952 under the leadership of Max Perutz (Received Nobel Prize for the structure of Haemoglobin) this institution has produced 9 Nobel Prize winners.

Protein Finger printing- Peptide Mapping

This technique involves the generation and 2-D analysis of peptides from a protein. Each protein has a unique peptide map (2-D analysis) and hence serves as a fingerprint for the protein. The steps involved in generating a peptide map/fingerprint are as follows **(Fig. 6)**.

1. Pure Hb and scHb are taken separately into test tubes.
2. The Hb and scHb are digested with the proteolytic enzyme trypsin which cleaves the protein after basic amino acid residues Arg and Lys.
3. Two separate strips of Whatman filter paper are spotted with Hb and scHb tryptic peptides and the peptides allowed to separate using the technique of paper electrophoresis at pH 2.0. Highly charged peptides will migrate more towards the anode/cathode.
4. The paper strips are dried, attached to larger squares of Whatman paper and chromatographed at right angles to the electrophoretic direction using a solvent system Butanol: Water:Acetic acid. In such a system peptides will separate based on their partition coefficient between the solvent and paper which is dependant on the relative hydrophobicity of the peptides. More hydrophobic peptides will move with the solvent to longer distances.

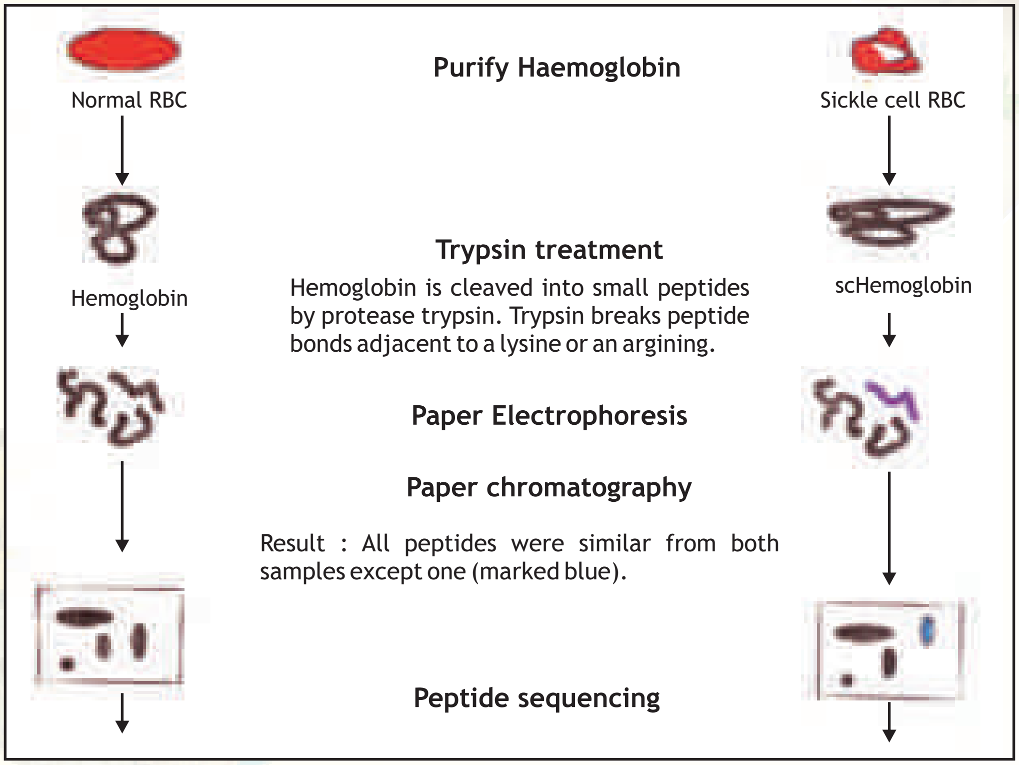
5. The chromatograms are dried and stained with a suitable visualisation reagent like Ninhydrin wherein peptide containing regions appear as orange yellow spots.

6. The peptide map for Hb and scHb are compared and it was found that one peptide was differently placed in the scHb map.

7. On examining this peptide and determining its amino acid sequence, Ingram found that it had a valine substitution for glutamic acid in the peptide.

   The single substitution of valine for glutamic acid (val/glu are at the 6th position of the haemoglobin beta chain) dramatically changes the structure of scHb making it form fibres within the RBC resulting in the deformation of the cell (sickling). Since the disease was due to a molecular alteration the term molecular disease was applied.

Peptide mapping became a useful technique to compare similar proteins from different sources. Slowly the information became too vast and computers were used to store this data into databases so that homology searches could be made. The protein fingerprinting data has been further augmented with new databases containing 2-D electrophoresis patterns of entire proteins from a given cell type, a technique developed by O'Farrel.



Hemoglobin is cleaved into small peptides by protease trypsin. Trypsin breaks peptide bonds adjacent to a lysine or an argining.

Result : All peptides were similar from both samples except one (marked blue).

Fig. 6. Protein fingerprinting

2-D Gel Electrophoresis

Two different techniques are combined in this procedure- Isoelectro focussing (IEF) and SDS-PAGE

(Fig. 7).

In simple electrophoresis, the mobility of proteins is due to their charge, which is pH dependant. At its isoelectric pH (pI), a protein does not possess any charge and thus will not move in an applied electric field. This feature is exploited in the technique of IEF, which separates proteins on the basis of their different pI values. Usually IEF is performed in thin tube gels. A pH gradient is set up within the IEF gel by the inclusion of polymeric buffers known as ampholytes. These, like proteins have many positive and negative charges and hence varying pIs. Because of their smaller sizes they move rapidly in an electrophoretic run setting up pH gradients when they come to rest at specific distances from the anode/cathode when they have no net charge. A protein sample from a cell or any other source is then electrophoresed within these tubes wherein the different proteins separate and migrate to their pI zones. The tubes containing the separated proteins is then laid on a SDS-PAGE slab gel and electrophoresis continued at right angles to the IEF direction.

In SDS-PAGE proteins separate on the basis of their size and hence at the end of this electrophoretic run proteins are separated into 2-D patterns with high resolution as two properties of the proteins have been exploited in their separation- charge and size. Proteins in the gels are stained with silver stains or other highly sensitive dyes and can be scanned, and pictures stored into computer databases for analysis.

5.2.4. Purification of Proteins

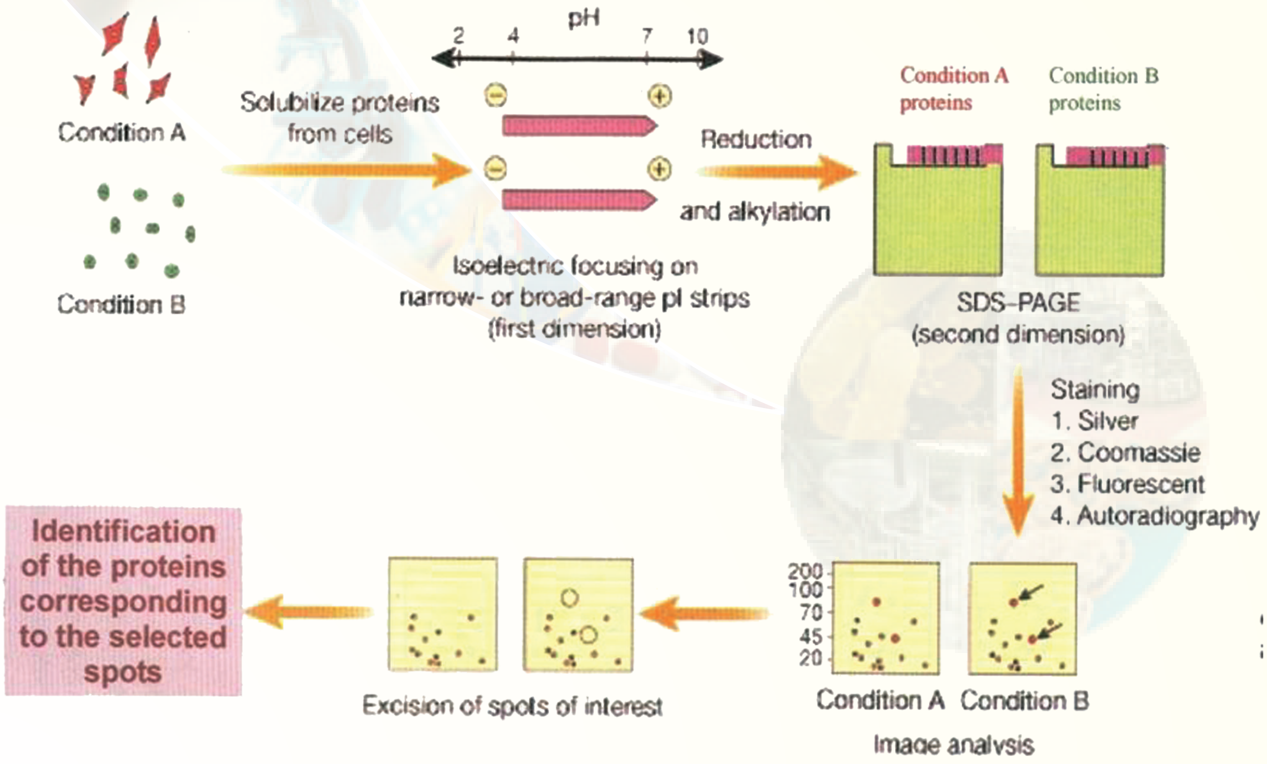

Condition A
Condition B
Solubilize proteins from cells
pH
2
4
7
10
⊖
⊕
Isoelectric focusing on narrow- or broad-range pI strips (first dimension)
Reduction
and alkylation
Condition A proteins
Condition B proteins
SDS-PAGE (second dimension)
Staining
1. Silver
2. Coomassie
3. Fluorescent
4. Autoradiography
200
100
70
45
20
Condition A
Condition B
Image analysis
Excision of spots of interest
Identification of the proteins corresponding to the selected spots

Fig. 7. Two dimentional gel electrophoresis

Isolation of a protein from a microbial culture, plant and animal sources involves various separation techniques (refer Textbook of Biotechnology for Class XI). These steps are collectively known as **downstream processing**. In spite of a large biodiversity of microbes we are restricted to certain bacteria/organisms which can be used as a source of protein as well as for introducing genes. These microorganisms are designated as "generally regarded as safe" (GRAS). **GRAS listed organisms are non-pathogenic, non-toxic and generally should not produce any antibiotics.** Similarily, plant tissue derived enzymes which have application in the food industry must be obtained from only non-toxic, edible plant species. One of the best known industrially useful enzymes is **papain** obtained from the latex of the green fruit and leaves of the Papaya tree. This enzyme finds application in meat tenderisation, clarification of beverages, digestive aids and wound cleaning solutions.

The existence of slaughter house facilities in which large number of animals are regularly processed to provide meat has also facilitated the collection of significant quantities of a particular tissue required as a protein source. Insulin is a classic example of a peptide hormone obtained from the pancreas of cows and pigs till the 1980's. Classical biotechnology required the slaughtering of over 100 pigs or 15 cows to meet the insulin requirements for one diabetic person for about one year. A rough estimate puts the number of diabetics in India by the year 2020 as 20 million! Obviously the requirement of insulin cannot be met by slaughtering pigs and cows. Fortunately, the advent of genetic engineering has ensured the availability of recombinant human insulin expressed in bacteria. Attempts are on to create transgenic animals by direct micro-injection of DNA into ova or stem cells and produce insulin and other proteins in milk on a commercial scale. This technology is called Molecular **Pharming (**Producing pharmaceuticals using genetically modified plants or animals). Advantages of producing recombinant proteins in milk are:

1. High production capacity.
2. Ease of source material collection (milking cows).
3. Moderate capital instrument requirements and low operational cost.
4. Ease of production including purification and scale-up.

Some medically useful peptides such as oxytocin have also been produced by direct chemical synthesis. In spite of different sources of proteins, the general principles of purification are similar. The overall approach and techniques are outlined in Unit-II, Chapter-III of the class XI, Textbook of Biotechnology. The exact details of the purification scheme for any given protein will depend upon a number of factors such as:

1. Exact source material chosen and location of the target protein (intracellular or extracellular).

2. Quantity of protein required and hence amount of raw material processed.
3. Physical, chemical and biological properties of the protein.

Calculation of amount of bacterial ferment required

Question: An *E. coli* cell produces at least 2000 different proteins. One of these is our enzyme of interest produced at a level of 3000 molecules per cell under optimum conditions. If we have to purify 1g of this intra-cellular enzyme, estimate how many cells of bacteria will be required theoretically? It is given that the molecular weight of the enzyme of interest is 1,00,000.

Answer: 1,00,000 g of the protein of interest corresponds to 1 mole of enzyme which corresponds to 6.023 x 10^{23} molecules (Avagadro no.)

Hence 1 g of enzyme has 1/1,00,000 x 6.023 x 10^{23} or 6.023 x 10^{18} molecules .

3000 molecules of the enzyme are present in one cell.

Therefore, 6.023 x 10^{18} molecules are present in 6.023 x 10^{18}/3000 = 2.007 x 10^{15} cells.

Question: Assuming that the bacterial cell is a cylinder (d = 1μm, h= 2μm) calculate (a) the total packed cell volume of *E.coli* required to produce 1 g of intra-cellular enzyme (b) the volume of the fermentor required if the maximum cell concentration inside the fermentor is 5% (cells need space to multiply).

Answer: Volume of a single bacterium = πr2h (cylinder volume) = 3.142 x 0.5 x 0.5 x10^{-12} x2 x10^{-6}

(note 1μm = 10^{-6}m, r = ½d) =1.57 x 10^{-18} m^3

2.007 x10^{15} cells (see previous answer) would have a volume of 2.007 x 10^{15} x 1.57 x 10^{-18} m^3

= 3.15 x 10^{-3} m^3 = 3.15 L Answer (a)

(1L = 10^{-3} m^3)

Answer (b) 100% concentration = 3.15 L

Therefore 5% concentration = 100/5 x 3.15 = 63 L

Volume of the fermentor required would be more than 63 L (30% extra space) about 82 L.

The source material chosen will dictate the range and type of contaminants present in the starting material. If the protein is extracellular, then one needs to separate the cellular components and process the medium to isolate the protein of interest. However if the protein is intracellular then the choice of method of cell disruption will depend on the cell type. Plant and fungal cells require harsher breakage methods; animal cells are easier to break because of no cell wall. Bacterial cells being very small require high pressure techniques. Once the proteins are released into suitable buffered solutions a variety of physico-chemical techniques are applied to selectively purify the protein of interest from the others.

Genetically engineered proteins are often tagged with certain molecules in order to confer some very pronounced physico-chemical characteristics on the protein of interest. This renders its separation from contaminants more straightforward. The ability to detect and quantify the total protein levels is an essential pre-requisite to the purification and characterisation of any protein. A typical purification scheme can be analysed as follows **(Table 1)**.

Table 1. Typical purification table

Procedure	Total protein (mg)	Activity (units)	Specific activity units/mg
Crude extract	20,000	40,00,000	200
Precipitation (salt)	5,000	30,00,000	600
Precipitation (pH)	1,000	10,00,000	1000
Ion-exchange chromatography	200	8,00,000	4000
Affinity chromatography	50	7,50,000	15,000
Size exclusion chromatography	45	6,75,000	15,000

Note that the last column is a good indication of whether a purification step is useful or not. This is because as a protein is purified its specific activity increases because the denominator should ideally decrease as irrelevant proteins are removed and only the specific protein/enzyme of interest is concentrated. Hence from the given table it is apparent that the step following affinity chromatography, size exclusion chromatography, is redundant as the specific activity does not change. For activity measurements it is also important to choose a proper assay method reflecting the sensitivity required and further the method should be specific. In the case of proteins absorbance measurements at 280 nm is easy, fast and non-destructive procedure for monitoring the concentration.

Bioassays can sometimes be more sensitive than chemical assays. However one needs suitable standards of known bioactivity values to arrive at a correct activity of the unknown sample. Where a protein of biological interest is concerned, example insulin, bioassays are mandatory. If the sample particularily has to be injected other safety tests such as toxicity have to be performed.

Downstream Processing

After cells (bacterial, animal or plant) have grown to their requisite capacity in a fermentor it becomes necessary to harvest the cells or medium depending in which component the recombinant protein is expressed and then purify the protein from other substances. These processes are part of downstream processing and because of the large amounts of source (a fermentor can be more than 1000 L capacity) bulk separation methods are used which are different from laboratory scale purification although principles involved are similar.

In the case of intracellular microbial proteins, cell harvesting is done by filtration or centrifugation from the fermentation medium, followed by re-suspension of cells in buffer or water with subsequent cell disruption. Most proteins obtained from plant and animal tissues are intracellular in nature. The initial step involves collection of the appropriate tissue, for example collection of blood to obtain proteins, collection of pituitary glands to obtain pituitary hormones etc.

Aqueous two-phase partition

When a crude cell homogenate is added to a biphasic mixture of dextran and polyethylene glycol (PEG) the cellular debris partitions to the lower, more polar and dense phase, dextran. Separation of the two phases achieves effective separation of cellular debris from soluble protein **(Fig. 8)**.

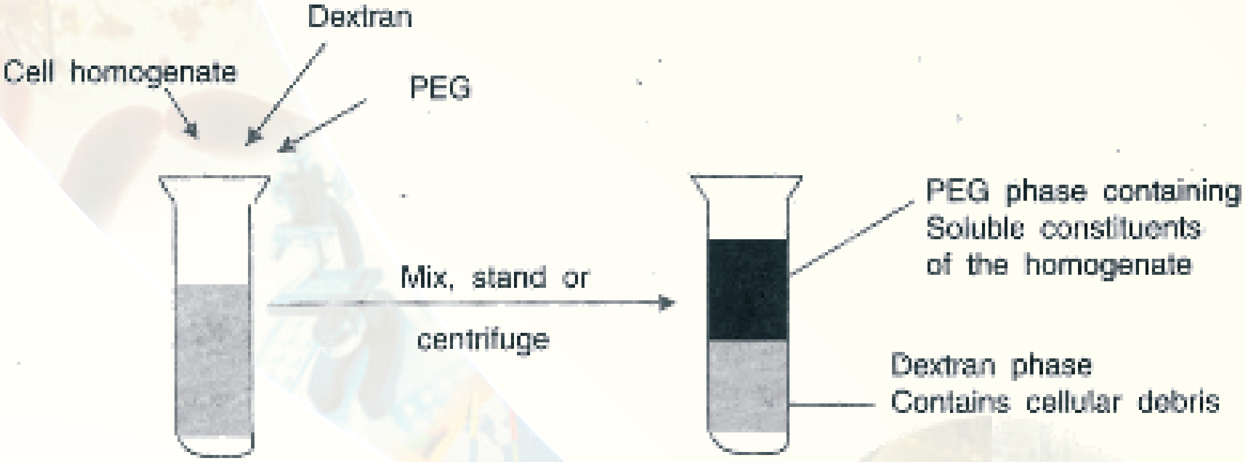


Fig. 8. Two phase separation

In some cases it is desirable and necessary to remove or destroy the lipids and nucleic acid of a cell homogenate as it may be a contaminant and can interfere with subsequent purification steps. The lipid layer can be removed by passage of solution through glass wool or cloth of very fine mesh size. Effective removal of nucleic acids may be achieved by precipitation or by treatment with nucleases.

For large scale application, concentration of extracts is normally achieved by precipitations, ion-exchange chromatography or ultrafiltration.

At any given pH value proteins display either a net positive, negative or no charge. Using these parameters different protein molecules can be separated from one another by judicious choice of pH, ionic strength and ion-exchange materials.

All efforts should be made to maximise protein stability during various steps. Some of the general conditions which may be followed are:

1. Maintenance of a specific pH value range of buffered solutions in which a protein is maximally stable.
2. Maintenance of physiological conditions (%CO_2 for animal cell culture and temperature).
3. Use of inhibitors to prevent the action of proteolytic enzymes.
4. Avoidance of agitation or addition of chemicals which may denature the target protein.
5. Minimise processing time.

Industrial scale production of proteins

The laboratory scale design cannot be scaled up to industrial scale directly. The following points need attention for industrial scale production:

1. Bulk purchase of chemicals and other raw materials would bring down costs.
2. The labour cost decreases sharply with increase in production.

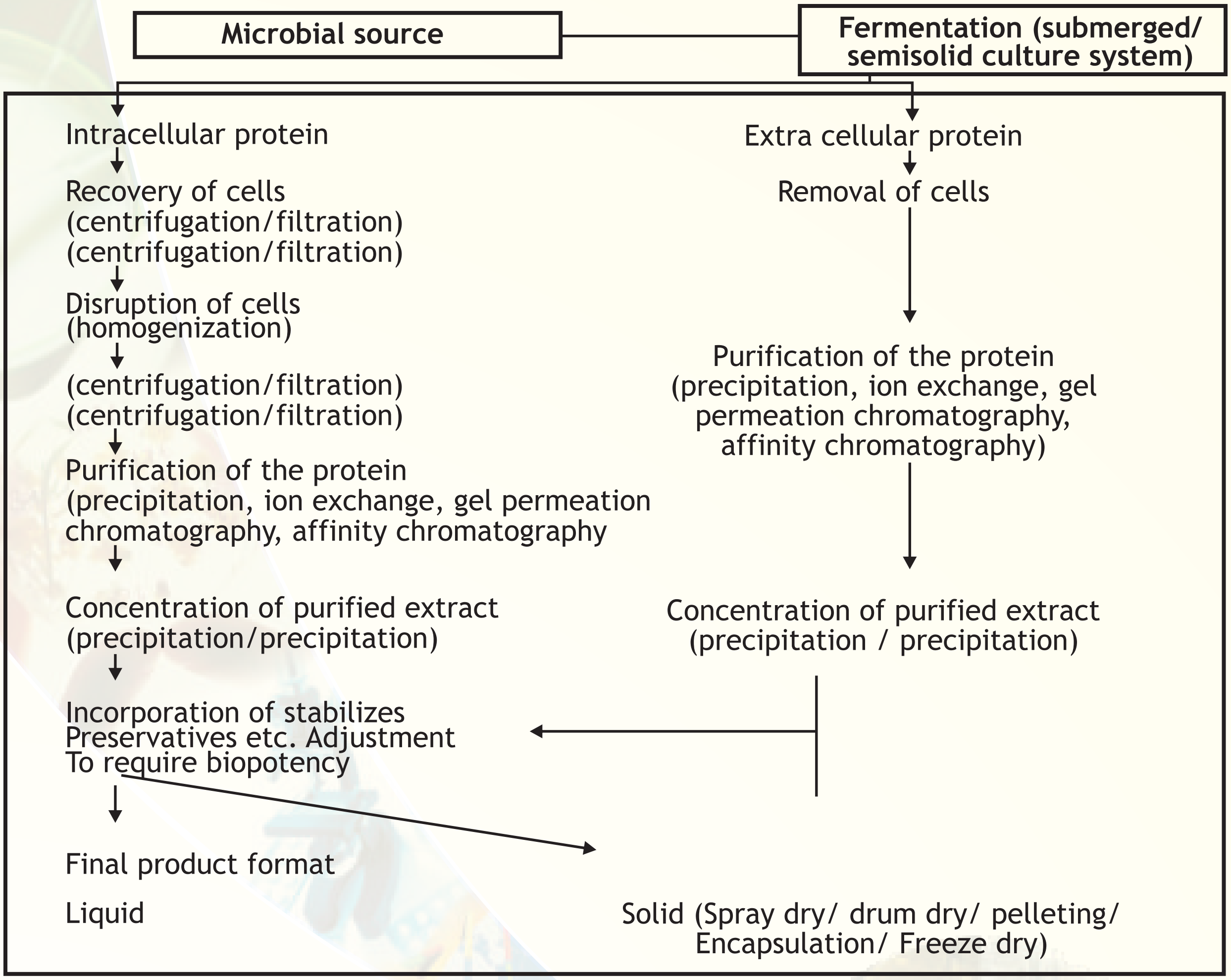
Most large scale process equipment such as holding vessels and transfer pumps are constructed from stainless steel or plastics, such as polypropylene. Glass vessels, so commonly used in laboratory scale culturing techniques are seldom used for large scale preparatory work. Materials used for large scale culturing must be inert and resistant to the corrosive action of any chemical used during the process. They should not allow any leaching of potentially toxic metals or chemicals into the product stream. It is useful to remember that any commercial plant has to have good GMP (good manufacturing practice) . Any downstream processing requires the approval of a regulatory authority in the form of a license to produce and market proteins designed for use in the food or health care industry. This ensures that the processing procedures are based upon established, validated methodologies. Generally 80% of the overall production cost are due to steps in downstream processing and quality assurance.

A generalised downstream processing scheme used in the production of bulk protein/enzyme from microbial sources is given in **Fig. 9.** Similar steps can be applied for animal and plant sources.

Special techniques for therapeutic /diagnostic proteins

These proteins must be purified to a very high degree especially for use in parenteral (injectable) administration. They also have to be sterile products (free of bacterial and fungal contamination) and hence can be administered by injection, infusion or implantation.

Fig. 9. A typical flow sheet from source to product.


Microbial source
Fermentation (submerged/ semisolid culture system)
Intracellular protein
Recovery of cells (centrifugation/filtration) (centrifugation/filtration)
Disruption of cells (homogenization)
(centrifugation/filtration) (centrifugation/filtration)
Purification of the protein (precipitation, ion exchange, gel permeation chromatography, affinity chromatography
Concentration of purified extract (precipitation/precipitation)
Incorporation of stabilizes Preservatives etc. Adjustment To require biopotency
Final product format
Liquid
Extra cellular protein
Removal of cells
Purification of the protein (precipitation, ion exchange, gel permeation chromatography, affinity chromatography)
Concentration of purified extract (precipitation / precipitation)
Solid (Spray dry/ drum dry/ pelleting/ Encapsulation/ Freeze dry)

5.2.5. Characterisation of Proteins

Techniques listed below characterise proteins with respect to properties such as mass, isoelectric charge, amino-acid sequence etc. Alongside they can also detect impurities as these are very sensitive techniques requiring small amounts, often micrograms of the sample. The first four techniques have already been discussed in this chapter.

1. Electrophoretic techniques, SDS/PAGE.
2. Fingerprinting.
3. Two dimensional gel electrophoresis.
4. Protein sequencing.
5. Mass spectrometry.

Mass spectrometry

Mass spectrometry (MS) has emerged as an important tool in biotechnology. It is extremely useful in obtaining protein structural information such as peptide mass or amino acid sequences. It is also useful in identifying the type and location of amino acid modification within proteins. One of the major attractions of mass spectrometry is that as little as picomoles (10^{-12}) of a protein sample can be analysed. A mass spectrometer is an analytical device that determines the molecular weight of chemical compounds by separating molecular ions according to their mass/charge ratio (m/z) ratios. The molecular ions are generated either by a loss or gain of a charge (e.g. electron ejection, protonation or deprotonation). After the ions are formed they can be separated according to their m/z ratio and finally detected. The process of ionisation, ion separation and detection in a mass spectrum can provide molecular weight or even structural information. A sample M with a molecular weight greater than 1200 D can give rise to multiple charged ions such as (M+nH)n+. Proteins/peptides have many suitable sites for protonation as all the backbone amide nitrogen atoms could be protonated theoretically as well as certain amino acid side chains such as lysine and arginine which contain primary amine functional groups.

A schematic diagram of the various parts of a mass spectrometer is indicated in **Fig. 10.** Basically a vapourised sample of a protein or peptide is introduced into the instrument wherein it undergoes ionisation. The charged molecules are then electrostatically propelled into a mass analyser (filter) which separates the ions according to their m/z ratio. The signal received upon detection of the ions at the detector is transferred to a computer which stores and processes the information.

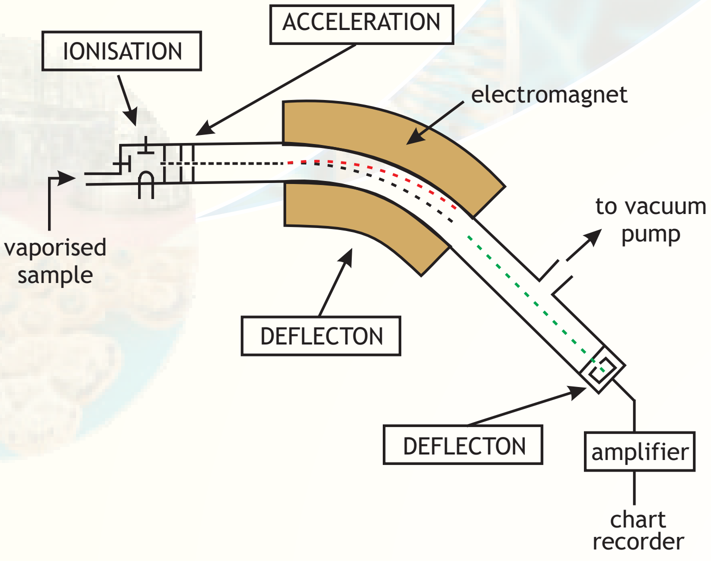


Fig. 10. An outline of a mass spectrometer.

The goal of mass spectrometric analysis of biomolecules like peptides and proteins is to create gas phase ions from polar charged molecules which are generally non-volatile. A popular method called Matrix Assisted Laser Desorption Ionisation (MALDI) is used to volatalise and protonate peptides and proteins. In this procedure, the sample is transferred from a condensed phase to a gas phase with the help of a solid matrix. Ion formation in MALDI is achieved by directing a pulsed laser beam onto a sample suspended or dissolved in a matrix. The matrix plays a key role in this technique by absorbing the laser light energy and causing the matrix material to vaporise. In the gas phase, the matrix plays a role in sample ionisation. The charged molecules are directed by electrostatic lenses from the ionisation source to the mass analyzer.

Multiply charged ions on analysis show patterns as indicated in Fig. from which molecular mass can be deduced as also indicated in the legend below **Fig. 11.**

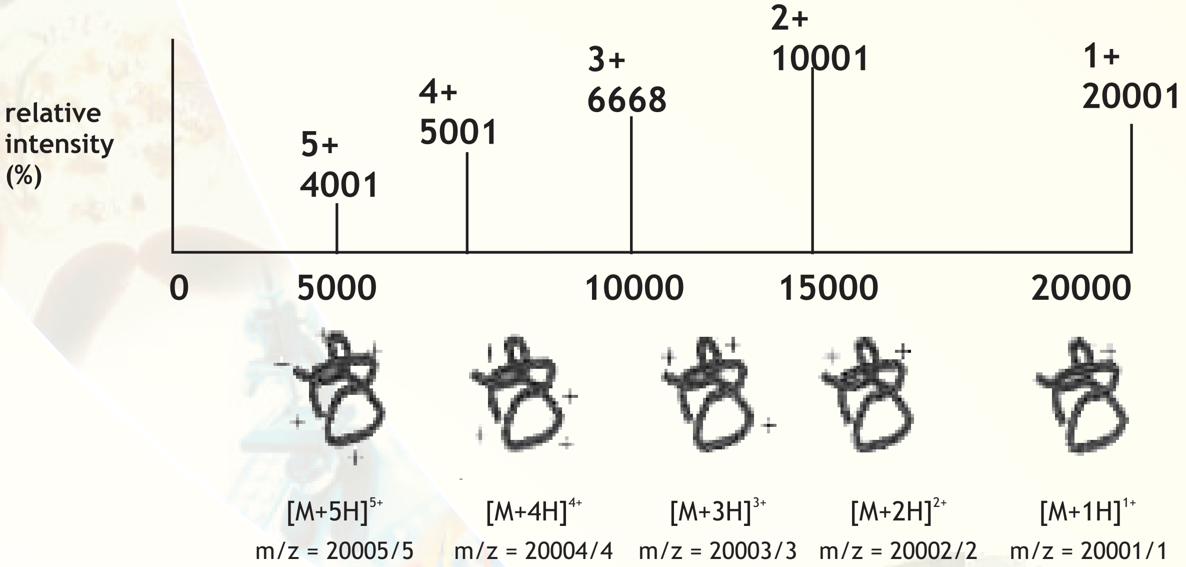


Fig.11. The same protein with a molecular weight of 10,000 contains 5, 4, 3, 2 and 1 charges. The mass spectrometer detects the protein ions at m/z = 2001, 2501, 3334, 5001 and 10,001 respectively.

Typically for protein identification, a crude extract is separated on a 2D gel. After visualising the proteins on the gel, protein spots are excised and used for mass analysis by the MS technique described above. For doing so the protein in the 2D gel after extraction from gel is either used intact or it is cleaved into small peptides with a protease like trypsin which makes the mass analysis easier. These peptides are separated on an on-line liquid chromatography system before introduction into the mass spectrometer. Liquid chromatography techniques like ion exchange, affinity or reverse phase column chromatography can be used to separate the peptides. These peptides are either sequenced directly or the mass of peptides is analysed using database searches (see Bioinformatics unit). With the human genome sequencing project, it is now possible to identify new proteins by combining the mass spectrometric information with the genomic information using Bioinformatics tools.

5.2.6. Protein Based Products

From the commercial point of view, proteins may be classified into the following categories.

1. Blood products and vaccines.
2. Therapeutic antibodies and enzymes.
3. Therapeutic hormones and growth factors.
4. Regulatory factors.
5. Analytical application.
6. Industrial enzymes.
7. Functional non-catalytic proteins.
8. Nutraceutical proteins.

Blood products and vaccines

Blood carries out several functions and is one of the best mediums for transportation in an animal. A better understanding of haematapoiesis (formation of blood cells) as well as factors responsible for blood coagulation has led to the discovery of several useful proteins. Several proteins from blood and plasma have been commercially available for decades. While these products have traditionally been obtained from blood donated by human volunteers, some are now produced by recombinant DNA technology. For example Factor VIII for treatment of Haemophilia A, Factor IX for treatment of Haemophlia B, Hepatitis B vaccine for prevention of hepatitis etc.

Therapeutic antibodies and enzymes.

Polyclonal antibodies have been used for more than a century for therapeutic purposes. More recently monoclonal antibody preparations as well as antibody fragments produced by recombinant DNA technology have found therapeutic use. For example tissue plasminogen activator (t-PA) is a proteolytic enzyme used to digest blocks in arteries following myocardial infarctions. A monoclonal antibody OKT-3 is used to prevent rejection following kidney transplantation because the antibody blocks those immune cells which attack foreign grafts.

Therapeutic hormones and growth factors

A number of hormone preparations have been used clinically for many decades. Though insulin was prepared from the pancreas of cows and pigs, the ability to genetically transfer human insulin gene into bacteria and the ability to modify amino acid residues (protein engineering) has facilitated the development of modified forms which are faster acting like humulin. Humulin acts in 15 min unlike pig insulin which takes 3 hours. Another growth factor- platelet derived growth factor has been approved for diabetics who develop skin ulcers. Several other growth factors are under various stages of clinical trials.

Regulatory factors

Several new regulatory factors were discovered that did not fit the classical definition of a hormone. Initially they were known as cytokines. These include interferons, interleukins, tumor necrosis factor and colony stimulating factors. The interferon family of INF alpha, beta and gamma have found widespread therapeutic application; interferon alpha is used for treatment of Hepatitis C, beta for Multiple Sclerosis and gamma for Chronic Granulomatous disease.

Analytical applications

Enzymes and antibodies have found a range of analytical applications in the diagnosis of diseases; hexokinase for quantitative estimation of glucose in serum, uricase for uric acid levels in serum, horse radish peroxidise and alkaline phosphates in ELISA etc.

Industrial enzymes

Proteolytic enzymes constitute an 8000 crore annual market for industrial enzymes. They find application in the beverage industry, detergent industry, bread and confectionary industry, cheese production, leather processing and meat industry. Alcalase is an enzyme used in the soap industry, papain is used in the beverage industry, glucose isomerise in the confectionary industry and chymosin is used in the cheese industry.

Functional non-catalytic proteins

Functional non-catalytic proteins are those which have properties such as emulsification, gelation, water binding, whipping and foaming etc. **(Table 2)**. For example kappa casein, a component of casein is involved in micelle stabilisation of milk proteins and keep the proteins suspended uniformly in milk because it behaves like a lipid molecule (2/3rd of the protein is hydrophobic). The food industry has exploited these non-catalytic proteins as illustrated in the below.

Table for whey protein.

TABLE : 2

Functional Property	Mode of action	Food System
Whipping/Foaming	Forms stable film	Egg less cakes, desserts, whipped topping
Emulsification	Formation and stabilization of fat emulsions	Vegetarian sausages, salad dressings, coffee whiteners, soups, cakes, infant food formulas, biscuits.
Gelation	Protein matrix formation and setting	Meat, baked goods, cheeses
Viscosity	Thickening, water binding	Soups, gravies, salad dressings
Water binding	Hydrogen bonding of water; entrapment of water	Meats, sausages, cakes, breads
Solubility	Protein solvation	Beverages
Browning	Undergoes Maillard reaction (on heating, the amino groups of protein react with aldehyde groups of sugars)	Breads, biscuits, confections, sauces
Flavour/Aroma	Lactose reacts with milk proteins	Baked goods, biscuits, confectionaries, sauces, soups, dairy products.

Nutraceutical Proteins

Nutraceutical is a word coined from combination of nutrition and pharmaceuticals. It has been observed that several nutritional proteins also have therapeutic functions. For example whey protein concentrates, lactose free milk (for lactose intolerant babies) and infant food formulations.

Where does one get the raw building materials such as amino acids needed to make all body proteins? During infancy we depend on milk. Baby milk formulations are also there (Amul, Lactogen etc.) which have been formulated to have similar composition as mother's milk. All these food materials provide the essential components nutritionally for growth and development during the first few months of our existence. A typical composition of milk from buffalo, human and cow sources is given in the **Table 3** from which baby milk formulations can be made to suit an infant.

Table 3. Composition of milk from buffalo, human and cow.

Constituents (per 100 ml of milk)	buffalo	human	cow
1. Protein (g)	3.8	1.2	3.3
2. Casein (g)	3.0	1.4	2.8
3. Lactalbumin (g)	0.4	0.3	0.4
4. Lactoglobulin (g)	0.2	0.2	0.2
5. Fat (g)	7.5	3.8	3.7
6. Lactose (g)	4.4	7.0	4.8
7. Calorific value (K Cal)	100.0	71.0	69.0
8. Calcium (mg)	203.0	33.0	125.0
9. Phosphorous (mg)	130.0	15.0	96.0
10. Chloride (mg)	112.0	43.0	103.0

From the **Table 3** it can be observed that milk contains several proteins, carbohydrates, lipids, vitamins, antibodies, minerals etc. It is interesting to note that human milk has nearly half the amount of casein as compared to cow and buffalo. Besides use of milk as a nutritional source, claims have been made to the effect that curd is beneficial in the management of some types of intestinal infections according to our ancient Sanskrit scriptures dating back to 6000 BC. Since time immemorial whey (liquid part of curds) has been administered to the sick for the treatment of numerous ailments. In 1603, Baricelli reported on the therapeutic use of cow's or goat’s whey, sometimes mixed with honey and herbs. The spectrum of illnesses treated with whey include jaundice, infected skin lesions, genitor-urinary tract infections. Gallen and Hippocrates insisted on a minimum daily drinking of one litre of whey. Using modern scientific research it has been possible to explain these observations. Whey proteins result in the elevation of a tripeptide glutathione(gamma-glutamyl cysteinyl glycine) in cells. This peptide is a reducing compound and has a broad range of functions including detoxification of xenobiotics and protection of cellular components from the effect of oxygen intermediates and free radicals. More recently curd has also been used as a pro-biotic (administered with antibiotics) because it is a good source of beneficial bacteria which can colonise the intestinal tract. **Table 4** gives the useful components of whey.

Table 4. Whey components.

α - Lactalbumin	Lactose
α - Lactoglobulin	Sialic Acid
Bovine Serum Albumin	Lactic Acid
Immunoglobulins	
Lactoferrin	Sodium
Lactoperoxidase	Potassium
	Calcium
	Magnesium
Protease peptones/polypeptides	Chloride
Free amino acids	Phospphate
Urea	Sulfate
	Citrate
Glycomacropeptides	Heavy Metals
Growth Factors	Milk Fat
	Globule
	Free Fat
	Lipoproteins

5.2.7. Designing Proteins (Protein Engineering)

Considerable interest exists in the biotechnology industry for the engineering of proteins with increased stability when exposed to harsh conditions like elevated temperature, organic solvents and reactive chemicals, often encountered in the industrial processes. Besides, it is of great interest to explore biological adaptations to environmental stresses such as high salinity, drought, cold etc. Therefore, in order to stabilise your favourite protein, it is essential to know the cause of inactivation.

Stability in a folded protein is a balance between the stabilising (mainly hydrophobic) interactions and the tendency towards destabilisation caused by the loss of conformational entropy as the protein adopts the unfolded form. The stability of a protein may however be changed by substituting amino acids that either favour stabilising interactions in a folded protein or destabilising interactions in an inactive protein. Numerous attempts have been made on different

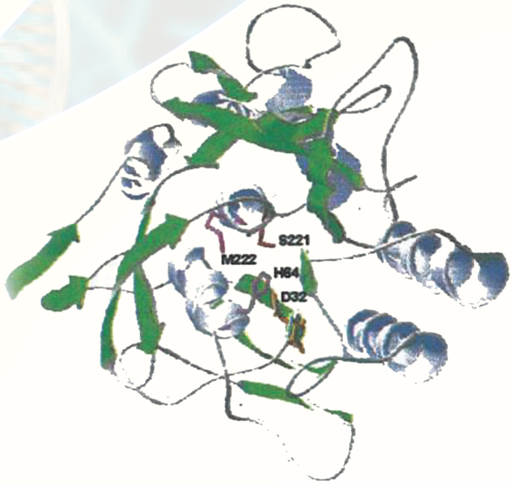


Figure 12. Subtilisin with the catalytic triad

proteins and enzymes in order to improve their properties for thermal and pH stability, solvent tolerance and solubility, catalytic potency etc. Given below is an example which has been successfully used in the detergent industry.

Improving laundry detergent Subtilisin

Subtilisin (27 kD) is a protease produced by bacteria that can digest a broad range of proteins that commonly soil clothing, see **Fig. 12**. The enzymatic activity of subtilisin is contributed by a catalytic triad, i.e., Ser221, His64 and Asp32 similar to chymotrypsin. Replacement of all three residues with alanine either singly or in combination results in significant loss of activity. Subtilisin represents the largest industrial market for any enzyme. To improve the efficiency of laundry detergents, detergent manufacturers supplement subtilisin in their products with various catchy slogans on the detergent box such as "stain cutter" or "biologically active enzymes".

The native enzyme subtilisin is easily inactivated by bleach (up to 90%). Careful studies showed that this inactivation was due to oxidation of the amino acid residue Methionine222 in the protein molecule **(Fig.12)**. Using site-directed mutagenesis of the subtilisin gene in *E.coli*, this methionine was substituted by a variety of other amino acids and the enzyme activity measured in the presence of bleach **(Table 5)**. It was observed that substitution of Met222 with Ala222 was the best in terms of activity and stability. Nowadays, many laundry detergents contain cloned, genetically engineered or recombinant subtilisin.

Table 5. Site-directed mutagenesis at codon position 222.

Codon-222	% activity w.r.t wild type	Codon-222	% activity w.r.t. wild type
Cys	138.0	Gln	7.2
Met	100.0	Phe	4.9
Ala	53.0	Trp	4.8
Ser	35.0	Asp	4.1
Gly	30.0	Tyr	4.0
Thr	28.0	His	4.0
Asn	15.0	Glu	3.6
Pro	13.0	Lie	2.2
Leu	12.0	Arg	0.5
Val	9.3	Lys	0.3

Creation of Novel Proteins

Conventional vaccines have utilised heat inactivated bacteria/viruses or their surface proteins to generate immunity against various specific diseases. Often it has been observed that some of the components of the vaccines have undesirable effects such as fever and rarely one has also heard of some of the components like virus actually causing the disease due to incomplete inactivation. Since proteins are the main molecules which provide the stimulus for immunity, attempts have been made to engineer proteins having minimum deleterious effects. The specific sequences of amino acids in the protein which stimulate immune response are known as epitopes. A recombinant vaccine based on selected epitopes may provide optimal design, scope for micromanipulation, unhindered supply and safety needed for an effective vaccine. Working on these lines, a novel synthetic gene has been assembled as a first step towards developing a subunit vaccine against Hepatitis B virus.

Improving nutritional value of cereals and legumes

The cereal grains and seeds of legumes constitute a major chunk of dietary protein requirement. The seed storage proteins are synthesised and accumulated throughout seed development to serve as source of amino acid reserves at the time of seed germination. High levels of such proteins in seeds would provide an enriched amino acid source for human consumption. However deficiencies in seeds of certain essential amino acids render the cereal grains or legumes unsuitable for a balanced diet. Supplementation of diet with essential amino acids from other sources therefore becomes essential. **Fig.13** gives the essential amino acid content of various cereals and commonly eaten food proteins. Essential amino acids are those which have to be obtained from food and cannot be made in our cells. From the data in the figure it is apparent that whey protein is superior to other sources especially with regard to branched amino acids- ile, leu, val, lys and trp. The branched chain amino acids (BCAA) are essential for the biosynthesis of muscle proteins. They help in increasing the bio-availability of high complex carbohydrates intake and are absorbed by muscle cells for anabolic muscle building activity. One of the theories is that during exercise the BCAAs are released from the skeletal muscle; the carbon skeleton part is used as fuel and the nitrogen part is used to make alanine which then goes to the liver where it is turned into glucose for energy. So for athletes who want to protect their existing mass, the idea is to take BCAA enriched foods before and after excercise. BCAAs reduce muscle breakdown and act as an energy source before and after exercise. Hence while maintaining exercise performance and delaying exhaustion BCAAs are very important for muscle growth. Nowadays an entire new area of sports medicine and nutrition prepare and recommend special nutrient drinks etc. which incorporate these principles. In the unit on plant tissue culture you will read how plant cereals have been genetically engineered for higher nutrient value in terms of proteins, vitamins etc.

Biological value (BV) measures the amount of protein nitrogen that is retained by the body from a given amount of protein nitrogen that has been consumed. It has been observed that the BV of whey proteins is the highest compared to rice, wheat, soya and egg proteins. Another index of protein value is the **protein efficiency ratio** (PER). PER is used as a measure of growth expressed in terms of weight gain of an adult by consuming 1g of food protein. The PER value of the following proteins are arranged in decreasing order- whey, milk, casein, soya, rice, wheat. The modern day approach for overcoming the nutritional deficiencies of seeds would be to engineer genes that would encode storage proteins with more of the nutritionally desirable amino acids either by inserting additional amino acids or substituting existing amino acids with new ones. Attempts are already being made on zein storage protein genes of maize to enhance its nutritional value. Introduction of entirely novel proteins that are highly enriched in specific amino acids is also being considered.

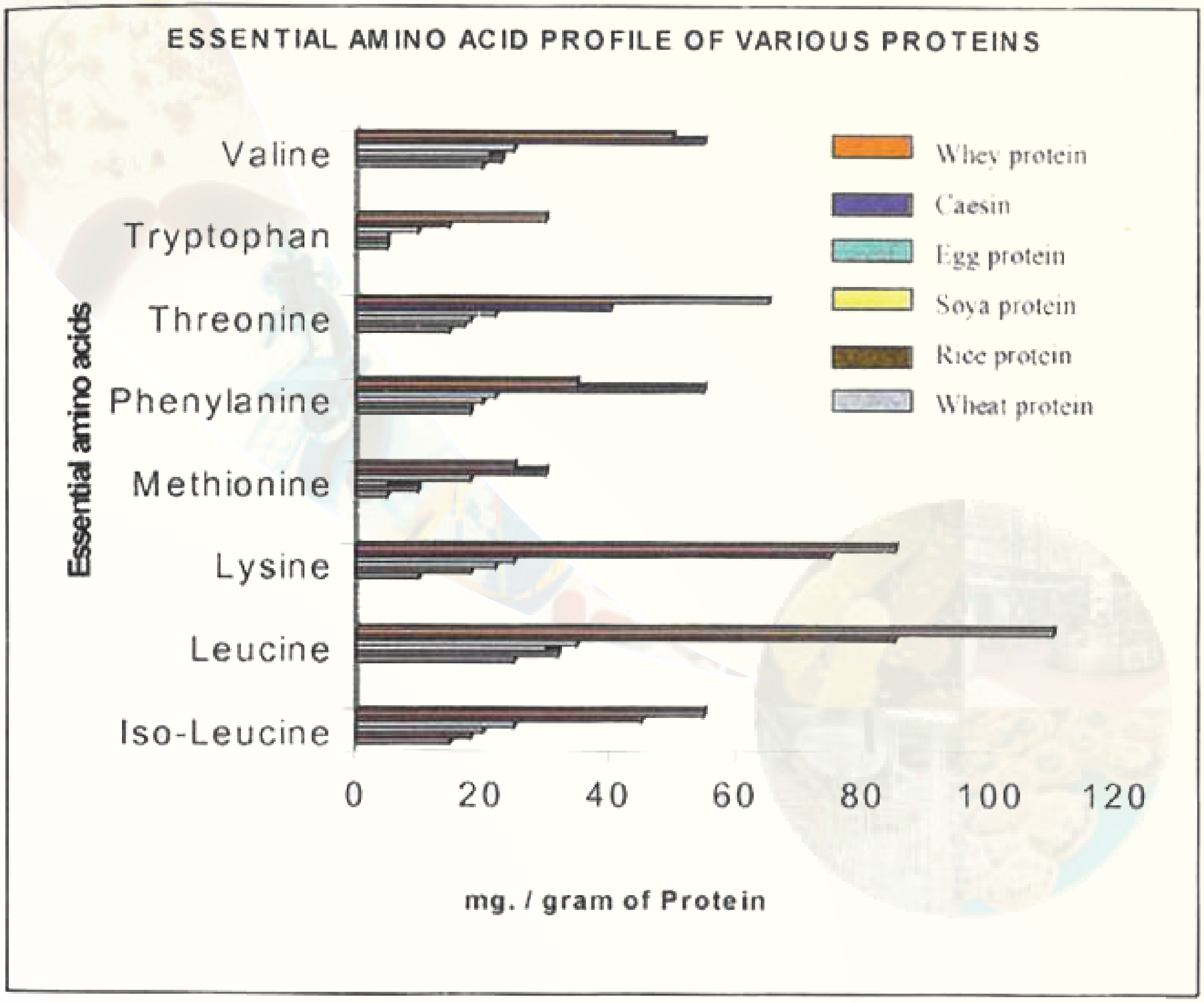


Fig. 13. The essential amino acids profile.

Review Questions

1. Name two human diseases caused by the absence of a protein.

2. What is the consequence if a protein is incorrectly folded? Give an example to illustrate your answer.

3. Distinguish between chymotrypsinogen and chymotrypsin.

4. Briefly explain how the serine residue in some enzymes can become acidic (reactive). Also suggest how you can confirm that a serine residue is involved in the catalysis.

5. Why is sickle cell anaemia called "Molecular disease"? How can sickle cell haemoglobin be identified?

6. What are the principles behind Isoelectric Focusing and SDS/PAGE techniques? Why is 2-D electrophoresis better than single dimension electrophoresis?

7. Define subunit, domain, and quaternary structure in proteins.

8. With an example explain the development of one protein based product.

9. What are non-catalytic functional proteins, therapeutic proteins and nutraceutical proteins? Give one example each.

10. Briefly discuss the use of designing a protein for any product.

11. What is the principle of MALDI-TOF? What is its main use in protein studies?

12. *E. coli* is a rod shaped bacteria about 2 μm long and 1 μm in diameter. The average density of a cell is 1 .28 g/ ml. Approximately 13.5% of the wet weight of E. coli is soluble protein. Estimate the number of molecules of a particular enzyme per cell if the enzyme has a molecular weight of 100,000 and represents 0.1 % of the total soluble protein. (Answer: 1626 molecules per cell).

References

1. Protein: Biochemistry and Biotechnology by Gary Walsh (2002), John Wiley & Sons, Ltd.

2. Practical Biochemistry: Principles and Techniques Edited by Keith Wilson and John Walker (2000), Cambridge University Press.

3. Biochemistry by Lubert Stryer, J.M, Berg and J.L. Tymoczko (2002). W.H. Freeman.

4. Principles of Biochemistry by Albert L. Lehninger, D.L. Nelson and M.M. Cox (2000), W.H. Freeman.

5. Biochemical calculations by Irwin H. Segel (1985), John Wiley & Sons, Ltd.

6. Proteins: Structures and Molecular Properties by Thomas E. Creighton (1992), W.H. Freeman.

CHAPTER 3 GENOMICS AND BIOINFORMATICS

5.3.1. Introduction

The term "GENOMICS" was coined in 1986 by Thomas Roder, to describe the scientific discipline of mapping, sequencing and analyzing genomes. H. Winkler in 1920 had coined the term genome to implicate the complete set of chromosomal and extra chromosomal genes of an organism, a cell, an organelle or a virus.

The field of genomics relies upon bioinformatics, which is the management and analysis of biological information stored in databases. During the mid-1980s to late 1980s, researchers started to use computers as central sequence repository, from where the data could be accessed remotely. Later in the early 1990s, genomics was transformed from an academic undertaking to a significant commercial endeavor, a course followed by bioinformatics a few years later.

In retrospect, the genomics really began with the conception of the Human Genome (HGP) in the mid-1980s. In the United States, the Human Genome Project officially started on October 1, 1990, as a 15-year program to map and sequence the complete set of human chromosomes, as well as those of several model organisms. The goal of sequencing an estimated three billion base pairs of the human genome was ambitious, considering that few laboratories in 1990 had sequenced just 100, 000 nucleotides. By 1993, the Human Genome Project had become an established international effort. The strategy of this international project was to make a series of maps of each human chromosome at increasingly finer resolutions. In this approach, chromosomes were divided into smaller fragments that could be cloned and then the fragments were arranged to correspond to their locations on a chromosome. After mapping, each of these ordered fragments would be sequenced.

Progress in stages

J. Craig Venter, a researcher at the National Institutes of Health, and his colleagues~ in early 1990s devised a new way to find genes. Rather than taking the Human Genome Project strategy of sequencing chromosomal DNA-one base at a time, his group isolated messenger RNA molecules, copied these RNA molecules into DNA, and then sequenced a part of these DNA molecules to create expressed sequence tags, or "ESTs." These ESTs could be used as handles to isolate the entire gene. Venter's method, therefore, focused on the "active" portion of the genome, which was producing messenger RNA for protein synthesis. The EST approach has generated enormous sized databases of nucleotide sequences, and facilitated the construction of a preliminary transcript map of the human genome. The development of the EST technique is considered to have demonstrated the feasibility of high-throughput gene discovery (screening of all possible gene candidates from the EST library), as well as provided a key impetus for the growth of the genomics industry. After the success of these projects, Craig Venter moved again to sequence entire genomes.

Evolving approaches

He devised the "whole-genome shotgun strategy," which involves randomly breaking DNA into segments of various sizes and cloning the fragments into vectors. Since the fragments are randomly cleaved from the genome, they tend to overlap, and a genome assembly program is used to fit contiguous pieces by matching overlapping ends. This method was validated by sequencing the entire genomes of a few selected microorganisms.

This is how, the first set of whole genome sequences of the smallest genome Mycoplasma genitalium and Haemophilus influenzae Rd were released. To analyze the data, several computer programs had to be adapted to fast computers, Later several new programs were also written to accomplish the task of sequence assembly. Craig Venter established an organization called "The Institute of Genomics Research (TIGR)" located in Maryland U.S.A., and soon whole genome sequences were determined for many other bacteria including those that live in exotic environments such as hot temperature or deep sea vents. Several bacteria of medical importance were also sequenced. During this time, several groups from Europe also initiated whole genome sequencing of bacteria such as Mycobacterium tuberculosis and *Bacillus subtilis* at Pasteur Institute. Generally, in Europe large consortiums (group of organizations in various countries) were formed to complement each other's strengths.

The exciting commercial era of genomics began with the establishment of Celera Genomics that was dedicated to sequencing the human and the mouse genomes, compared to microbial genomes, the human genome is large ~ 3×10^9 bps and also contains lots of repeated sequences. These repeated sequences present difficulties in sequence assembly because doubts arise with regard to their true order of arrangement in the genome. The parts containing the genes were somewhat easier to assemble. The problem of sequence assembly of repeats and of unique sequences by the computer is akin to this example. Suppose you were blindfolded and asked to pick two balls of different size from a lot of mostly identical balls, you would make several attempts but end up with failures most of the time. Further, you may not be able to distinguish one ball from another. However, if you were given the same assignment of drawing two different balls from a lot balls of all different sizes and shapes, then there is a good possibility of you picking up two different balls at much fewer attempts, perhaps the very first attempt itself may be enough.

But the unraveling of the human genome sequence gave us a surprise. Initial EST sequencing had led to an estimate of over 100,000 genes being present in the human genome. The genome sequence however, revealed that there are only about 30,000 genes. This number is only twice that of the fruitfly *Drosophila melanogaster,* a simple organism compared to the immensely complex human being. Possessing only twice the number of genes of fruitfly challenges us to search for other explanations that underlie the complexity of the humans. It turns out that humans can achieve this through combinations of the genes. You can understand this by an example.

Suppose, a mechanic has a set of 20 or 30 tools, each one dedicated to carry out a specific task. Then the mechanic can accomplish 20 or 30 tasks. However, if the same set of tools had flexible parts, then the same mechanic can generate several 'new combinations' of these tools to carry out hundreds of tasks. Naturally through combinations, this mechanic will have more business, earn more money and will be most sought after compared to using the dedicated 'one job specific' tools.

Below, the sections on genomics explain the different branches of this exciting new area. With the development of automated sequencing, it has been possible to sequence genomes of many organisms. According to the latest list displayed at the NCBI (National Centre for Biotechnology Information) site, there are 1409 complete genome sequences of bacteria and archaea, 40 complete genome sequences of eukaryotes and 2537 complete genome sequences of viruses. The sequencing projects have shown several interesting and unexpected findings. The term genomics itself has undergone expansion in last few years and in the present context also includes genome function. Genomics can be broadly divided into structural genomics and functional genomics.

Structural Genomics

Structural genomics primarily involves high-throughput DNA sequencing followed by assembly, organization and management of DNA sequences. It represents an initial phase of genome analysis, which involves the construction of high-resolution genetic, physical or transcript maps of the organism. The ultimate physical map of an organism is its complete DNA sequence. Although, in the last few years with the completion of several genome-sequencing projects, the term structural genomics has also undergone transition. Several structural genomics initiatives now encompass systematic and high-throughput determination of three-dimensional structures of all proteins. The information and reagents provided by structural genomics are used to design global (genome-wide) experiments to identify functions of proteins.

Functional genomics

Functional genomics represents a new phase of genome analysis and deals with the reconstruction of the genome to determine the biological function of genes and the interactions between genes. The fundamental strategy in a functional genomics approach is to expand the scope of biological investigation from studying single genes or proteins to studying all genes or proteins at once in a systematic manner. Functional genomics is therefore characterized by high-throughput or large scale experimental methodologies combined with statistical and computational analysis of results.

5.3.2. Genome Sequencing Projects

There are several reasons for completely sequencing a genome.

- First it provides a means for the discovery of all the genes and thus provides an inventory of genes.
- Second, the sequence shows the relationships between genes.
- Third, it provides a set of tools for future experimentation.
- Fourth, sequencing provides an index to draw and organize all genetic information about the organism.
- Fifth, and very important over time, is that the whole genome sequence is an archive for the future containing all the genetic information required to make the organism.

There are several methods for small-scale sequencing, although most of these do not scale well to sequence entire genomes. The two main methodologies used for genome sequencing are discussed here. These have also been briefly discussed in the introduction.

Directed sequencing of Bacterial Artificial Chromosome (BAC) contigs

You have already learnt in the previous chapter that Bacterial Artificial Chromosome (BAC) vectors are capable of stably propagating large, complex DNA inserts in *Escherichia coli*. These vectors are used to make genomic libraries in which the insert size is 80-100 kb. This library is then screened by finding common restriction fragments. These BAC clones are then mapped to find overlapping arrays of contiguous clones called contigs. The mapped contigs are sequenced by breaking large DNA fragments into small pieces. Therefore, in this directed sequencing strategy, pieces of DNA from adjacent stretches of a chromosome are sequenced.

Random shotgun sequencing

Random shotgun sequencing is one approach to sequence genomic DNA. Genomic DNA macromolecules are very long and they contain many genes and other sequences required to build the whole organism. Even with the best of sequencing techniques we get a maximum of 700 bases of sequence information from one single run of an experiment. Therefore, we need a strategy to sequence the whole DNA. The random shotgun sequencing approach follows a very well known common theme "divide a big problem into small tasks. Solve these small tasks individually. Finally add up all these solutions to get the full final solution". Big genomic DNA molecules are broken down into small fragments, which are cloned in small (2.0 kb) and medium (10 kb) plasmid vectors. Plasmids have specific sites where these molecules can be inserted through enzymatic procedures. Thus, a library is constructed. Now each clone is picked up randomly and sequenced from both ends. By picking many clones and sequencing them, we get large amounts of sequences. Observations show that several of these sequences are identical, some are similar to each other in parts called overlapping parts, whereas, a few may be just unique. After we feed all these data into a computer program, these sequences are joined by finding overlapping parts. The result is, we get long pieces of DNA sequences. This process of assembling continues until all overlapping parts are exhausted. Finally, we would get a large portion of the genomic DNA sequence.

Even though in theory, the entire genomic DNA sequence can be obtained in this way, in practise, this is not so. Some gaps in genomic DNA sequence do arise and these gaps need to be closed by specific cloning of those regions and additional sequencing.

5.3.3. Gene prediction and counting

Gene prediction is an important problem for computational biology and there are various algorithms that do gene prediction using known genes as a training data set. Since most of the knowledge to carry out these predictions comes from experimentally identified genes, this becomes a limitation. Even if we know where the genes are in the genome, it is not entirely clear how to count them. Due to the existence of overlapping genes and splice variants it is difficult to define the parts of the DNA that should be regarded as the same or several different genes. Nevertheless, for practical purposes (allowing for some 'experimental error') we can count the number of genes in an organism. Some of the results of counting predicted genes have turned to be quite surprising **(Table 1)**.

Table 1. Genome size and gene predictions between several organisms.

Organism	No. of chromo somes	Genome size in base pairs	The Number of Predicted genes	Part of the genome that encodes for protein
Bacteria *Escherichia coli*	1	500,000	5000	90%
Yeast *Saccharomyces cerevisiae*	16	12,068,000	6340	70%
Worm *Caenorhabditis elegans*	6	100,000,000	19,000	27%
Fly *Drosophila melanogaster*	4	175,000,000 - 196,000,000	13,600	20%
Weed *Arabidopsis thaliana*	5	157,000,000	25,498	20%
Human *Homo sapiens*	23	3,000,000,000	20,000 - 25, 000	< 5%

One of the surprises is the relatively small number of genes in a human genome (20,000 - 25,000 genes) in comparison to worm (19,000 genes). In fact some experts still think that there must be at least 40,000 - 50,000 genes in the human genome, and that 30,000 just reflects the unreliability of in silico (i.e., computational) gene prediction. Still, it seems that there is no simple correlation between the intuitive complexity of an organism and the number of genes in its genome.

5.3.4. Genome Similarity, SNPs and Comparative Genomics

The human genome project has often raised two interesting questions - whose genome is being sequenced and how similar are genomes between two different individuals. It is understood that several anonymous samples were collected and pooled for human genome project. Since each person's genome is believed to be 99.8% identical to everyone else's, we can talk in terms of the consensus human genome. SNPs are DNA sequence variations, which occur when a single base (A, C, G, or T) is altered so that different individuals may have different bases at these positions **(Fig. 1).**

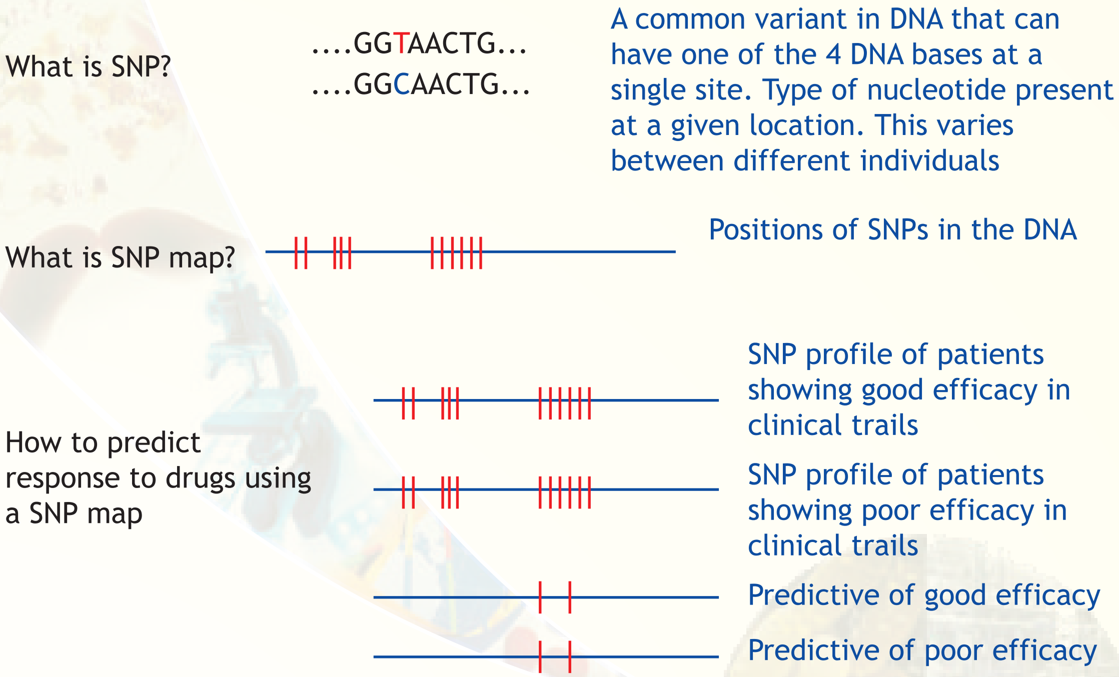

What is SNP?
....GGTAACTG...
....GGCAACTG...
A common variant in DNA that can have one of the 4 DNA bases at a single site. Type of nucleotide present at a given location. This varies between different individuals
What is SNP map?
Positions of SNPs in the DNA
How to predict response to drugs using a SNP map
SNP profile of patients showing good efficacy in clinical trails
SNP profile of patients showing poor efficacy in clinical trails
Predictive of good efficacy
Predictive of poor efficacy

Fig.1. Definition of SNP, and an illustration to show how physicians can use SNP map to determine how patients are likely to respond to a particular drug. The vertical bars indicate various SNPs on the human DNA.

However, the other way to look at it is that the 0.2% difference in DNA sequence is enough to make each individual unique. It is understood that on an average one in a thousand nucleotides are different in genomes of two different individuals. Particularly, important variations in individual genomes are the **single nucleotide polymorphisms** or SNPs, which can occur both in coding and non-coding regions of the genome. It is believed that SNPs occur at 1.6 million to 3.2 million sites in the human genome, and may affect gene function, depending upon exact base change and where it occurs. It would be interesting to note the following research based observation.

1. The genetic variations between individuals (particularly, in the non-coding parts of the genome) are exploited in DNA fingerprinting, which is used in forensic science.

2. However, not all genetic variations are beneficial **(see Table 2)**. Genomic variations underlie differences in our susceptibility to, or protection from all kinds of diseases. The severity of illness and the way our bodies respond to treatments are also manifestations of genetic variations. For example, a single base difference in the ApoE gene is associated with Alzheimer's disease, and that a simple deletion within the chemokine-receptor gene CCR5 leads to resistance to HIV (Human Immunodeficiency Virus) infections and the development of AIDS (Acquired Immunodeficiency Syndrome). SNP analysis is therefore important for diagnostics and a SNP database has been developed to aid these applications. An example of how a physician can decide if a medicine prescribed will be effective to a patient is illustrated in **Fig. 1** (vertical bars denote various SNPs in patients' genomes).

Table 2. Genes and diseases

Single-gene mutations which follow mendelian inheritance	**Gene polymorphisms which has complex inheritance**
Cystic Fibrosis (Cystic Fibrosis Transmembrane Conductance Regulator CFTR gene) 1. Inheritance: autosomal recessive disease	Common late-onset Alzheimer's disease 1. Inheritance: Major cause is epsilon4 allele of the gene coding for apolipoproteinE (APOE)
2. Genomic location: Chromosome 7 (7q31.2)	2. Genomic location: Chromosome 19 (19q13) and recently Chromosome 10 (10q21).
3. Mutation: The most common mutation is a deletion of 3 bps resulting in the loss of codon no. 508, which codes for phenylalanine	
Huntington disease (Huntingtin gene HTT) 1. Inheritance: autosomal dominant 2. Location: Chrosome 4 (4p16.3) 3. Mutation: increased number of CAG repeats more than 35 times	Migraine 1. Susceptibility locus: Chromosome 6p12.2 - 6p21.1 and Chromosome 1q31

3. Physicians can use patients DNA sample to determine the pattern of SNP genotype profile and from that they can predict how patients are likely to respond to a particular drug. SNP analysis can also be used in population genetics, as some SNPs vary in different frequencies between populations.

4. The genome sequencing projects have revealed that the genomes of organism otherwise quite different in appearance are quite similar for example mouse and man, are quite similar. Another example is that, among the conserved elements between different species such as the worm and the yeast, substantial portion belongs to genomic regions coding for proteins

5. It is estimated that the difference between human and chimpanzee genomes is only 1 to 3%, while human and mouse share about 97.5% of their working DNA. These similarities suggest that none of these genomes has changed much since we shared a common ancestor 100 million years ago.

5.3.5. Functional Genomics

Functional genomics dissects the emerging knowledge about genomes to understand the gene and their product functions and interactions. Two exciting new developments are now enabling scientists to get a wealth of clues to this complicated story. The new technique, microarray technology and proteomics provide snapshots of all the genes expressed in a cell or tissue under different environmental conditions. The DNA microarray technology is used for analysing the expression of thousands of messenger RNA molecules.

Fluorescence *in situ* hybridization

It is possible to introduce colours into DNA by a technique called Nick Translation developed in 1977 by Rigby and Paul Berg. The enzymes, DNA polymerase I makes DNA and DNase I, which cuts DNA are combined in a buffered reaction with dNTP's, including dUTP labelled with a red or green fluorescence. The DNA polymerase I adds nucleotide residues to the 3-prime hydroxyl terminus that is the result of nicks (breaks) created by the DNase I in the DNA. In the process, the fluorescence labelled nucleotide in the free nucleotide mixture becomes incorporated into the newly synthesized strands of DNA **(Fig. 2)** .

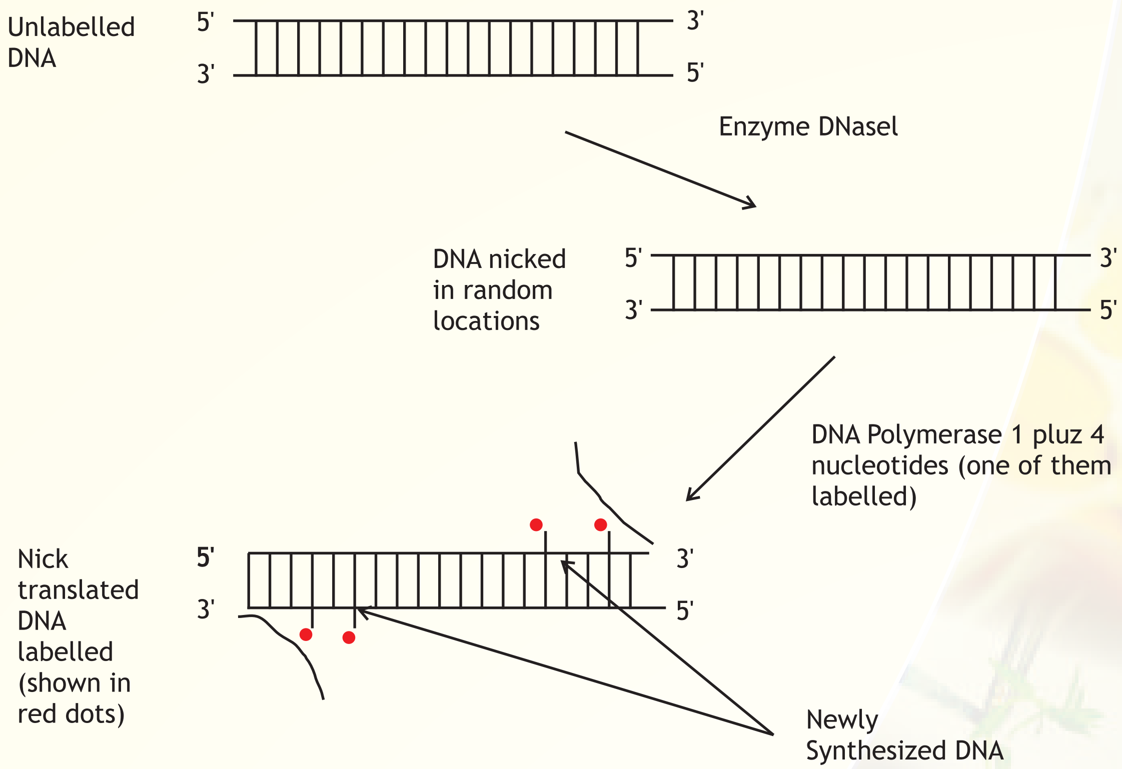

Unlabelled DNA
5'
3'
3'
5'
Enzyme DNaseI
DNA nicked in random locations
5'
3'
3'
5'
DNA Polymerase 1 pluz 4 nucleotides (one of them labelled)
Nick translated DNA labelled (shown in red dots)
5'
3'
3'
5'
Newly Synthesized DNA

Fig. 2. Nick Translation. Nicks (breaks) are created in the DNA using DNase I. Subsequently, DNA Polymerase I synthesizes new DNA using the older one as template and incorporates the labelled nucleotides in the process. Finally we get labelled DNA.

The DNA fragment size with fluorescence probe after nick translation depends upon the amount of enzyme and the incubation time of reaction. The size range can be 300 to 3000 bp. The application of FISH can be illustrated by taking an example of chronic mylogenous leukemia (CML). It was observed from the karyotype analysis of the lymphocyte preparation made from blood samples of CML patients that there was a 9-22 translocation in the chromosome (also called 'Philadelphia chromosome'). Although by counting the number of such cells it was possible to find out the severity of the disease, it was not an easy procedure. The regions on the chromosomes involved in translocation were identified on chromosomes 9 and 22. From the DNA library it was possible to pick up clones carrying the particular genes involved in CML. Using nick translation it was possible to flourescently label chromosome 9 region with red colour and chromosome 22 region with green colour and prepare the probe **(Fig. 3)**.

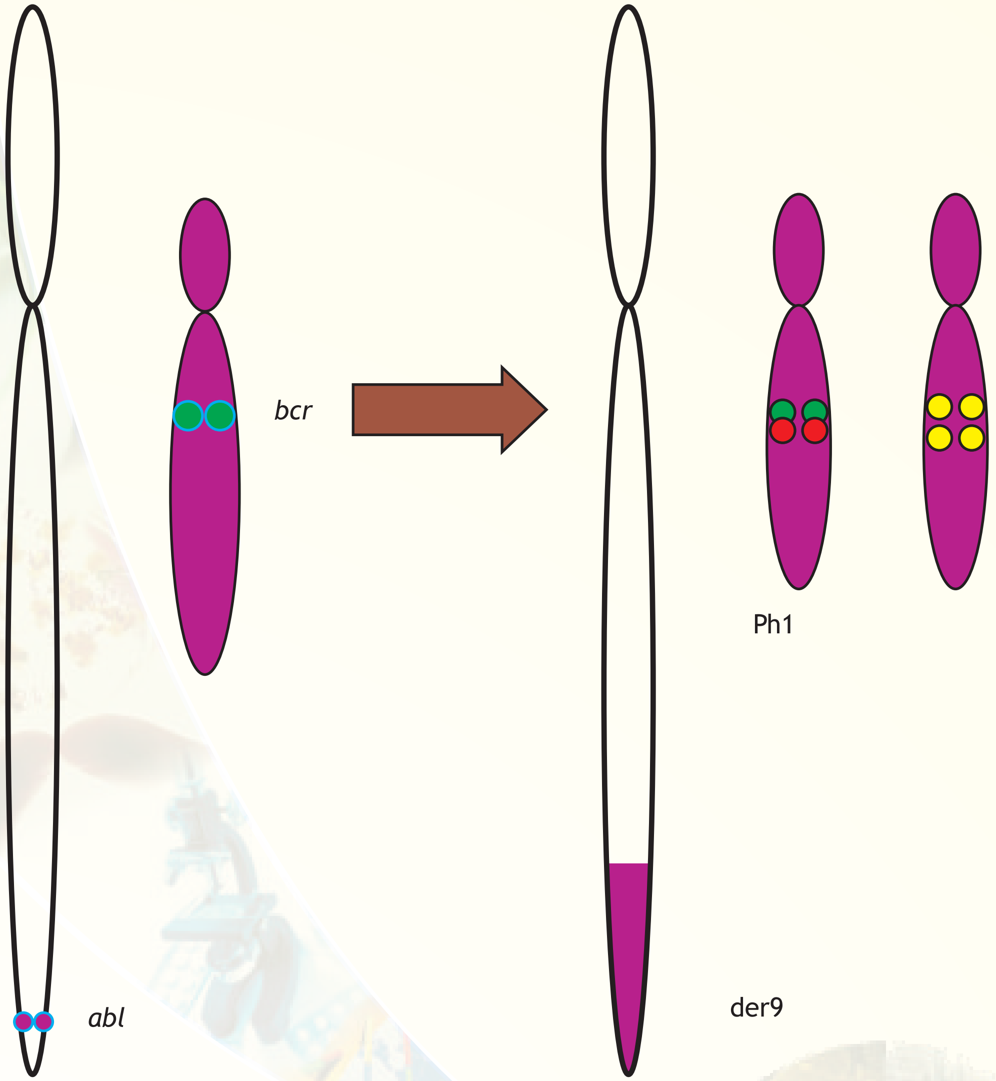


Fig. 3. Reciprocal translocation between Chromosome 9 and Chromosome 22 forms an extra-long chromosome 9 (der 9) and the Philadelphia chromosome (Ph1) containing the fused a-bcrg ene. This is a schematic view representing metaphase chromosomes.c

It was observed that when CML lymphocytes smear cells were hybridized with the two probes *in situ* and when observed under fluorescent microscope, the cells, which were affected, appeared yellow (mixing of green and red colour produces yellow colour). The unaffected cells appeared as red and green **(Fig. 3)**. This technique known as Fluorescence *in situ* Hybridization (FISH) allows knowing the status in the interphase unlike in karyotyping where you need a metaphase chromosome. The status of the disease could easily be identified by counting the number of cells, which appeared yellow. Further, it was possible to monitor the effect of chemotherapy and drugs by taking out samples and counting the number of cells appearing yellow.

Microarray Technology

It is widely believed that thousands of genes and their products (i.e., RNA and proteins) in a given organism function in a complicated and orchestrated way that creates the mystery of life. However, traditional methods in molecular biology generally work on a "one gene - one experiment" basis, which means that the throughput is very limited and the "whole picture" of gene function is hard to obtain. In the recent years, a new technology, called DNA microarray, has attracted tremendous interests among biologists. This technology promises to monitor the whole genome on a single chip so that researchers can have a better picture of the interactions among thousands of genes simultaneously.

Principle: Microarrays consist of large numbers of DNA molecules spotted in a systematic order on a solid substrate, usually a slide **(Fig. 4)**. The base pairing or hybridization is the underlying principle of DNA microarray. Microarray exploit the preferential binding of complementary single-stranded nucleic acids. A microarray is typically a glass (or some other material) slide, onto which DNA molecules are attached at fixed locations (spots). The type of molecule placed on the array units also varies according to circumstances. The most commonly used molecule is cDNA, or complementary DNA, which is derived from messenger RNA. Since cDNA are derived from a distinct messenger RNA, each feature represents an expressed gene. In order to detect cDNA bound to the microarray, they must be labeled with a reporter molecule that identifies their presence. This technique of introducing fluorescent dyes in DNA and its use in detection of target molecule by hybridization has been previously applied in fluorescent *in situ* hybridization (FISH).

Procedure: Comparative hybridization experiments compare the amounts of many different mRNA in two cell populations. If one wanted to compare a normal cell and a cancerous cell, the following experiments are needed to be carried out. mRNA is first purified from total cellular contents. mRNA accounts for only about 3% of all RNA in a cell so isolating it in sufficient quantity for an experiment (1-2 micrograms) can be a challenge. Since free RNA is quickly degraded, to prevent the experimental samples from being lost, they are reverse transcribed back into more stable DNA form. The products of this reaction are called complementary (cDNA) because their sequences are the complements of the original mRNA sequences. The reporters currently used in comparative hybridization to microarrays are fluorescent dyes (fluors), represented by the red and green circles attached to the cDNAs in **Fig. 4**. A differently-colored fluor is used for each sample so that we can tell the two samples apart on the array.

The labelled cDNA samples are called probes because they are used to probe the collection of spots on the array. The two cDNA probes are tested by hybridizing them to a DNA microarray. The array holds hundreds or thousands of spots, each of which contains a different DNA sequence. If a probe contains a cDNA whose sequence is complementary to the DNA on a given spot, that cDNA will hybridize to the spot, where it will be detectable by its fluorescence. In this way, every spot on an array is an independent assay for the presence of a different cDNA. Microarrays are made from a collection of purified DNA molecules typically using an arraying machine. The choice of DNA to be used in the spots on a microarray determines which genes can be detected in a comparative hybridization assay. The end product of a comparative hybridization experiment is a scanned array image. A small piece of such an image is shown in **Fig. 4**.

Interpretation: The measured intensities from the two fluorescent reporters have been false-coloured red and green and overlaid (in reality, you observe colours when the fluors are stimulated by a laser). Spots, whose mRNA is present at a higher level in one or the other cell population show up as predominantly red or green. Yellow spots have roughly equal amounts of bound cDNA from each cell population. Therefore yellow spots correspond to genes expressed approximately equally in both normal and cancerous cells whereas red spots correspond to genes expressed in high amounts in normal cells. Similarly green spots correspond to genes expressed in high amounts in cancerous cells. By drawing this distinction, we would be able to understand the altered gene expression patterns in cancerous cells. This allows us to further understand the mechanism and make attempts to develop cures.

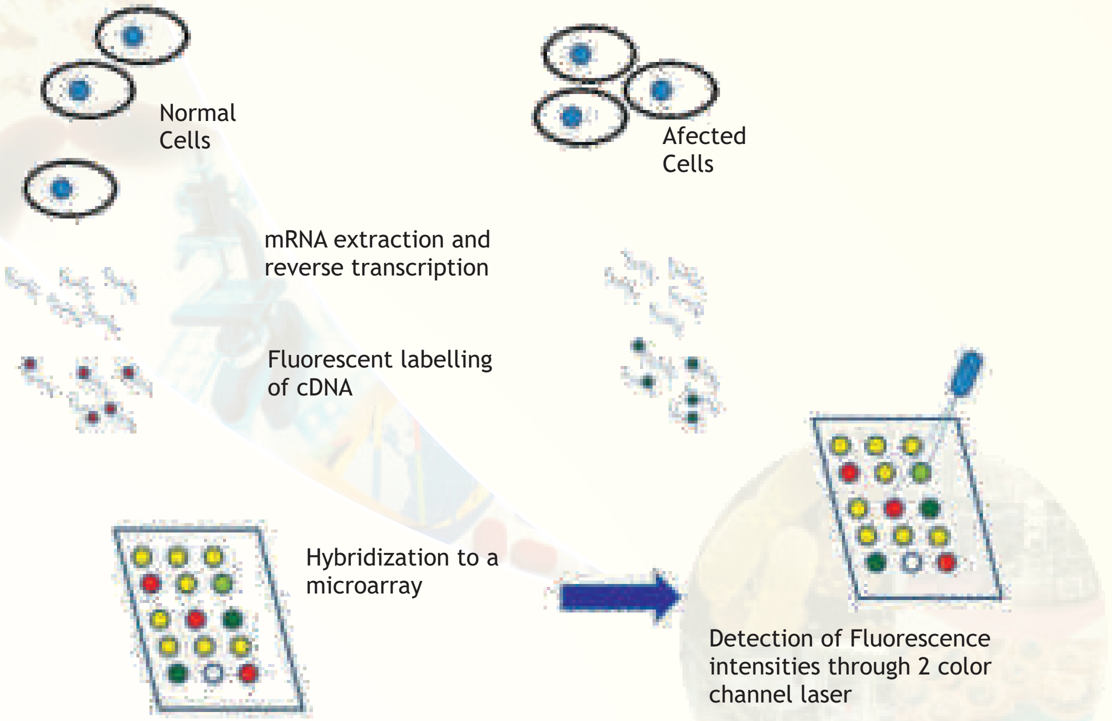

Normal Cells
Afected Cells
mRNA extraction and reverse transcription
Fluorescent labelling of cDNA
Hybridization to a microarray
Detection of Fluorescence intensities through 2 color channel laser

Fig. 4. Major steps involved in comparative microarray hybridization experiments between normal and affected (for example, cancerous) cells are illustrated. Observe that some spots are yellow, meaning that the particular gene is expressed in equal amounts. Some spots are clearly red or green indicating that the particular genes are expressed in only normal or affected condition. Some spots are more greenish or orange meaning that the expression status is not clearly tilted to either side but there is a trend towards either extremities. A few spots may appear blank (no colour).

This microarray technology promises to monitor the whole genome on a single chip so that researchers can have a better picture of the interactions among thousands of genes simultaneously. There are several names to this technology - DNA arrays, gene chips, biochips, DNA chips and gene arrays. In the case of gene chips, the substrate for immobilization is a silicon wafer and the probes are oligonucleotides spotted through photolithographic etching. In this case of gene chip only 1 colour hybridization is performed per chip. Comparisons are done by matching data from one chip to another through a special data normalization procedure. The principle used in this technology is being extended to develop protein arrays also. This technique has been used to study the following:

1. Tissue specific genes
2. Regulatory gene defects in a disease
3. Cellular responses to environment
4. Cell cycle variations

The strength of genomic studies lies in its global comparisons between biological systems. Genomics studies provide initial guidelines to identify areas for deeper investigation and to see how these results fit in the biological context.

5.3.6. Proteomics

The term Proteome refers to the complete protein set of a cell. Proteomics refers to the large scale characterization of the entire protein complement of cells, tissues and even whole organisms. Modern proteomic studies involve many different areas, which are illustrated in **Fig. 5**. These include protein-protein interaction studies, protein function, and protein localization. The growth of proteomics is a direct result of advances made in large-scale nucleotide sequencing of various genomes. Without this development, protein identification would have been difficult.

It is important to have information about the proteins simply because they are responsible for the phenotype of the cells. It is impossible to understand mechanisms of disease, ageing etc. solely by studying the genome. Only by understanding protein function and their modifications, drug-targets for various diseases can be identified. One of the major aims of proteomics is to create a three dimensional map of a cell indicating the location of proteins.

The proteome of a given cell is dynamic. In response to internal and external cues biochemical machinery of the cell could be modulated. This could lead to several changes in the proteins such as post-translational modifications, changes in cellular localization, effect on their synthesis or degradation. Thus examination of a proteome is like taking a snapshot of the protein environment at a given time.

It is speculated that no function can be assigned to about one-third of the gene sequences in organisms for which the genome sequence is known. The complete identification of proteins in a genome will help structural genomics projects. The aim of these projects is to obtain 3D structure of all proteins in a genome. The structural analysis would be helpful in assigning function to many of these proteins. In addition to identification of proteins, one of the major goals of proteomics is to characterize post-translational modifications on proteins.

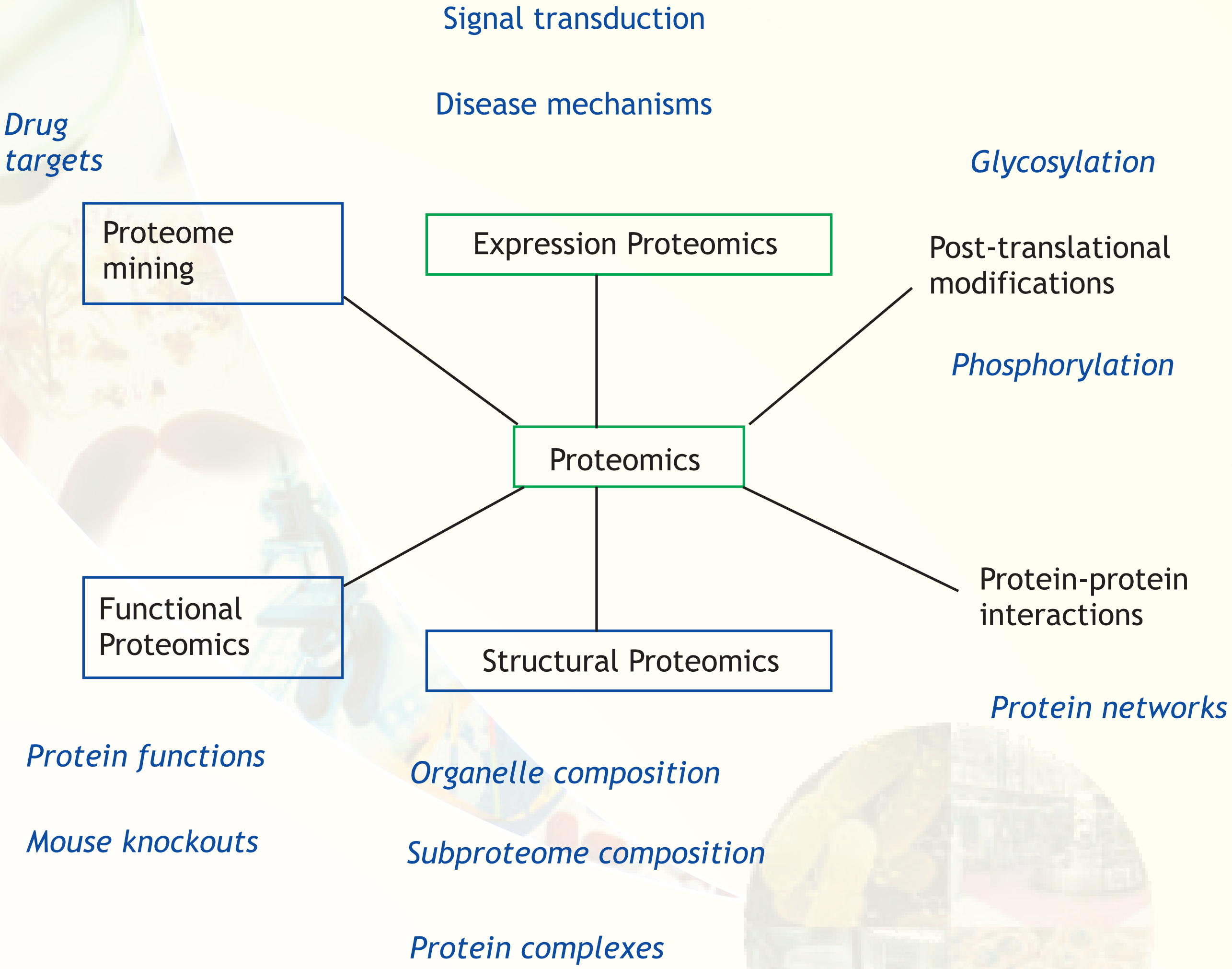


Fig. 5. Types of proteomics and the scientific knowledge that can be gained from them.

Types of Proteomics

Expression proteomics: The quantitative study of protein expression between samples that differ by some variable is known as expression proteomics. Using this approach, protein expression of the entire proteome or of subproteomes between samples can be compared. This could be useful in identification of disease specific proteins. For example: tumor samples from a cancer patient and a similar tissue sample from a normal individual could be analyzed for differential protein expression. Using two dimensional gel electrophoresis, followed by mass spectrometry, proteins, which are over or under expressed in the cancer patient compared to the normal individual can be identified. This could be compared with the microarray data **(Fig. 4)**. Identification of these could provide a lead in understanding the basis of tumor development.

Structural proteomics: Unlike comparing the same cell or tissue in normal and diseased state in expression proteomics, structural proteomics are directed to map out the structure and nature of protein complexes present specifically in a particular cellular organelle. The aim is to identify all proteins present in a complex and to characterize all protein-protein interactions occurring between these proteins. Isolation of specific sub cellular organelles or protein complexes by purification can help assembling information about architecture of cells and explain how expression of certain proteins gives a cell its unique characteristics.

Functional proteomics: Functional proteomics is a very broad term for many specific, directed proteomics approaches. It can be defined as the use of proteomics methods to analyze the properties of molecular networks involved in a living cell. One of the major objectives is to identify molecules that participate in these networks. One of the successes of functional proteomics is identification and analysis of protein networks involved in the nuclear pore complex. This discovery has led to the identification of novel proteins which are important for translocating important molecules from the cytoplasm of a cell to the nucleus and vice versa.

Genes and Proteins

Number of genes vs Number of proteins

Analysis of mRNA does not provide a direct reflection of the protein content in the cell. One reason is that mRNA and protein expression levels do not always correlate. The formation of mRNA is only the first step in a long sequence of events resulting in protein synthesis **(Fig. 6)**. mRNA could undergo various post-transcriptional modifications like, polyadenylation and mRNA editing. Some of these modifications could lead to the generation of various protein forms from a single gene. Subsequently, translational regulation of mRNA could take place. Proteins, after synthesis, could undergo post-translational modifications. It is estimated that proteins could undergo as many as 200 different types of these modifications. Due to these reasons the relationship between number of genes and number of proteins is not linear. In other words, the number of proteins could easily outnumber the number of genes.

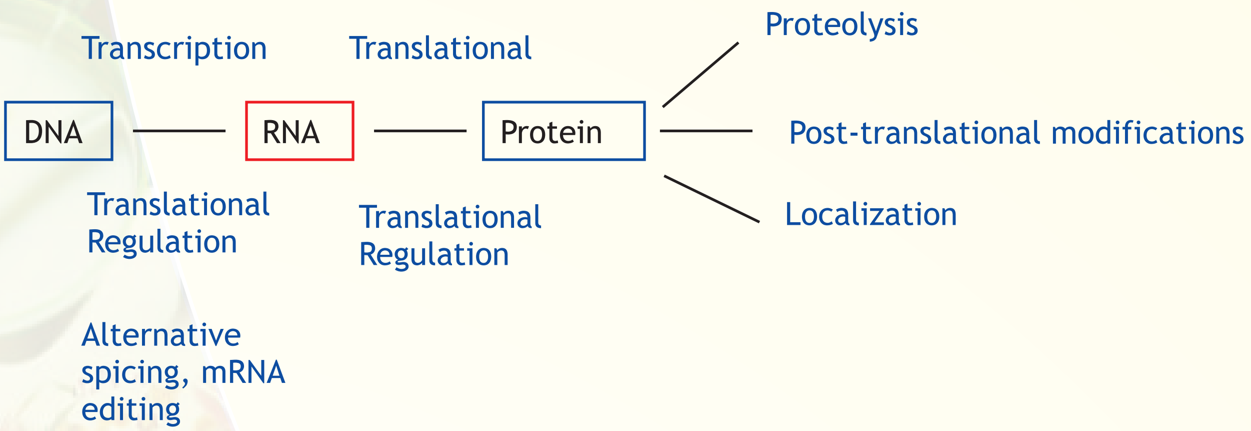


Fig. 6. Processes through which genes can give rise to multiple protein products with differing functions.

5.3.7. History of Bioinformatics

Bioinformatics has emerged as a scientific discipline that encompasses the application of computing science and technology to analyze and manage biological data. All this began when it was demonstrated by Ingram that there is homology between sickle cell haemoglobin and normal haemoglobin. This led to comparison of other proteins with similar biological function. As more and more proteins were sequenced, it became necessary to have databases which enabled a quick comparison using computational softwares. With the advent of rapid nucleic acid sequencing techniques, a large number of sequences started accumulating which again required computing facilities.

In 1962, Zuckerkandl and Pauling proposed a new approach of studying evolutionary relations using sequence variability. This initiated a new field called 'molecular evolution'. The approach was based on the observation that functionally related or homologous protein sequences were similar. Subsequently, sequence comparisons, analysis of functional relatedness and inference of evolutionary relationships became possible. Margaret Dayhoff observed that protein sequences undergo variation during evolution according to certain patterns. She noted that :

- amino acids were not replaced at random but were altered with specific preferences. For example, amino acids with similar physico-chemical characteristics were preferred, one for another.

- some amino acids such as tryptophan, was generally not replaced by any other amino acid.
- based on several homologous sequences, a point accepted mutation (PAM) matrix could be developed.

This laid the first foundation for subsequent work on sequence comparisons using quantitative approaches.

The National Biomedical Research Foundation (NBRF) compiled the first comprehensive collection of macromolecular sequences in the "Atlas of Protein Sequence and Structure' published from 1965-1978 under the editorship of Margaret O. Dayhoff. Dayhoff and her research group pioneered the development of computer methods for the comparison of protein sequences, for the detection of distantly related sequences and duplications within sequences, and for the inference of evolutionary histories from alignments of protein sequences.

In 1980, the data library was established at the European Molecular Biology Laboratory (EMBL) to collect, organize, and distribute nucleotide sequence, data and related information. Now its successor is the European Bioinformatics Institute (EBI) located at Hinxton, U.K. The National Centre for Biotechnology Information also started in USA as a primary information databank and provider at about the same time. Later, the DNA Data Bank of Japan was initiated. The Protein Information Resource (PIR) was established in 1984 by the National Biomedical Research Foundation (NBRF) as a resource to assist researchers in the identification and interpretation of protein sequence information. Today, all these databanks are in close collaboration with each other and they exchange data on a regular basis.

As the sequence data began to accumulate rapidly, new powerful sequence analysis softwares were needed. In parallel, firm mathematical basis was also required to develop algorithms. Scientists from the field of mathematics, biology, and computer science entered the emerging field of bioinformatics.

The databanks through their wide network of distribution of information are very important sources for all researchers who take interest in asking fundamental questions in biology. Thus, a major primary aim of bioinformatics is to spread scientifically investigated knowledge for the benefit of the research community. Other aims include the development of softwares for data analysis.

The word "bioinformatics" is a combination from biology and informatics. As it became clear that biological polymers, such as nucleic acid molecules and proteins, can be transformed into sequences of digital symbols informatics approaches can be used for analysis. Moreover, only limited set of letters is required to represent the nucleotide and amino acid monomers. It is the digital nature of this data that differentiates genetic data from many other types of biological data, and has allowed bioinformatics to flourish. Another key point is that the use of sequence data relies upon an underlying reductionist approach: sequence implies structure which in turn implies function. In the subsequent sections we will see the details of these activities.

5.3.8. Sequences and nomenclature

The nomenclature system we adopt in Bioinformatics work is based on the International Union of Pure and Applied Chemistry (IUPAC) recommendations. It is useful to follow this nomenclature system so that data sets from different laboratories situated around the world can be compared easily and uniformly. The database institutions and the journals that publish research reports follow these recommendations strictly to ensure uniformity and to aid rapid reproducibility. We will go through the basic nomenclature system for nucleic acids and proteins in this section. Details of modifications of the nucleotides may be touched upon but we suggest you refer to the IUPAC website for these details. For routine work using the nucleic acid and protein sequence data we discuss the following system of IUPAC nomenclature **(Fig. 7)**.

Human communication language	Biological language
Letters	Nucleotide bases
Words	Genes, Exons, Introns
Sentence	Operons
Punctuation	turns, kinks, bending
Chapter	Chromosome

Fig. 7. Comparison between Human communication and biological system. The biological language is used in Bioinformatics.

DNA and protein sequences

Table 3. Summary of single-letter code IUPAC recommendations.

Symbol	Meaning	Base(s)
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G Or A	puRine
Y	T Or C	pYrimidine
M	A Or C	Amino
K	G Or T	Keto
S	G Or C	Strong (3 Hydrogen bonds)
W	A Or T	Weak (2 Hydrogen bonds)
H	A Or C Or T	not-G, H follows G in the alphabet
B	G Or T Or C	not-A, B follows A in the alphabet
V	G Or C Or A	not-T (not-U), V follows U in alphabet
D	G Or A Or T	not-C, D follows C in the alphabet
N	G Or A Or T Or C	Any

The symbols, their meaning and the bases for the nucleic acid sequences are presented in **Table 3.** The first 4 bases G,A,T,C, their symbols and the basis for nomenclature is clear. While determining sequence data through experiments, sometimes, the sequence identity at a particular position may not be clearly identifiable due to compression artifacts or other secondary structure related problems. In most cases the problem can be solved by repeating the experiment and also by sequencing the complementary strand. In a few cases, ambiguities may persist. In such cases, the most probable results are inferred from the chromatograms.

For instance, at a position where the ambiguity is not resolvable between a 'G' or a 'C' but one can be sure that there is no possibility of "A' or 'T' in the same position, then the symbol to be used is 'S'.

In most organisms, DNA is present as double stranded. The two strands are anti-parallel and complementary to each other (following Watson-Crick base-pairing). However, the problem arises when we start encountering the symbols that mean more than one base at a given position. Again, the IUPAC system comes to aid. The symbols to be used in the complementary strand corresponding to the symbol at the same position in a given strand are specified in **Table 4**. In certain cases, the complementary symbols are same as in the given strand because in both cases they mean the same set of bases.

Table 4. Definition of complementary symbols.

Symbol	A	B	C	D	G	H	K	M	S	T	V	W	N	R	Y
Complement	T	V	G	H	C	D	M	K	S	A	B	W	N	Y	R

The symbols and their meaning for the protein sequences are presented in **Table 5.** It is evident that the number of symbols that mean more than one amino acid is very few.

Table 5. Symbol definitions for the amino acids.

Single letter code	Three letter code	Full name
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Aspartic acid
C	Cys	Cysteine
Q	Gln	Glutamine
E	Glu	Glutamic acid
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Methionine
F	Phe	Phenylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Threonine
W	Trp	Tryptophan
Y	Tyr	Tyrosine
V	Val	Valine
B	Asx	Asx
Z	Glx	Glx
X	Xaa	Xaa

The concept of directionality

In the biological systems, the usual direction in which the DNA and RNA are synthesized is the 5'-3' direction. This is universal and therefore it is helpful to adopt this fact as a way to collect and store data in the sequence databases. The nucleotide sequence are generally present in the database as they have been submitted or published, subject to some conventions which have been adopted for the database as a whole. The sequences are always listed in the direction 5' to 3'. Bases are numbered sequentially beginning with 1 at the 5' end of the sequence. The complementary sequence is described with a 'c' indicated next to the position of the sequence. Complementary sequence also runs 5'-3' but in the opposite direction to the given strand. Only one strand of the DNA sequence is given in a database entry. The complementary strand will have to be inferred using programs available in various packages or from various websites. We will see the details of a database entry in the next section. In the case of proteins, they are synthesized in the cell from N-terminus to the C-terminus. It is useful to adopt this convention in database entry for protein sequences. Thus, the concept of directionality used in biological systems is useful in describing the conventions to be adopted by the Database institutions. The advantage here is the universality of these fundamental biological processes in almost all living organisms.

Question : If you are given a sequence without any label, how will you find out whether it is a DNA sequence or a RNA sequence or a Protein sequence?

Answer : The usual approach taken by standard computer programs like sequence search programs scan the first 20 symbols. If the symbols encountered switch between any of the 4 bases only, then the sequence at hand is taken as a DNA sequence. Instead of T if U is encountered, then it is a RNA sequence. But if the symbols switch between any of the 20 (greater than 4), then it is taken as protein sequence.

Different types of sequences

cDNA : A large number of sequences deposited in the Databases were determined from cDNA molecules. While filling up the sequence entry form you must tick at the right position to indicate whether the sequence being deposited is a cDNA sequence. This data will also be provided when a sequence is retrieved. Thus in the case of cDNA sequences one is looking at the expressed part of the genome.

Genomic DNA: Sequencing of genomic DNA has become very routine nowadays. The genomic DNA is the store-house of information of which expressed part is represented in the cDNA sequences also.

ESTs : It is an abbreviation for Expressed Sequence Tags. Dr. Craig Venter initiated sequencing of a large number of cDNA molecules by sequencing one end of each of the randomly picked cDNA clones. Millions of ESTs have been deposited in a special database called dbEST. EST data is used to infer expression patterns by counting the number of ESTs corresponding to each gene divided by the total number of ESTs.

GSTs : In *Plasmodium falciparum* the enzyme Mung Bean Nuclease (MNase) cleaves in between the genes. A genomic DNA library generated by digestion with MNase was used for gene identification in *P. falciparum*. The approach used was similar to ESTs. One read of sequence was obtained from either ends. This data is referred to as genome sequence tags (GSTs). Usually, genomic DNA sequence refers to the nuclear DNA.

Organelle DNA: Eukaryotic cells have organelles such as mitochondria and chloroplast. These organelles have their own store house of information in the form of organelle DNA. Organelle DNA codes for a few genes. The coding information for the rest of the genes reside in the nuclear DNA of the same cell. If an organelle DNA has been sequenced the appropriate position in the sequence submission form must be mentioned.

Other molecules: In addition to these molecules, the databases contain the sequences of other molecules such as tRNA, and other small RNAs.

5.3.9. Information Sources

Major databases

The National Center for Biotechnology Information (NCBI): The NCBI at the National Institutes of Health was created in 1988 to develop information systems in molecular biology. In addition to maintaining the GenBank nucleic acid sequence database, NCBI provides data retrieval systems and computational resources for the analysis of GenBank data and the variety of other biological data made available through NCBI. Since then, the tools and the services provided by the NCBI have grown so large that it would require a full chapter to describe about NCBI alone.

The resources available from the NCBI have been classified into the following heads: (1) Database retrieval tools, (2) BLAST family of sequence similarity search programs, (3) Gene level sequences, (4) Chromosomal sequences, (5) Genome analysis, (6) Analysis of gene expression patterns, (7) Molecular structure.

All these Web based tools are available free. We shall learn about the practical aspects of some of these tools in practical classes. Below we discuss the first three set of resources. Learning these can help you in most cases. A great majority of bioinformatics activity is carried out using these resources. Other resources are useful for advanced studies.

Database retrieval tools

Among the database retrieval tools are ENTREZ, TAXONOMY BROWSER, LOCUS LINK. Entrez is an integrated database retrieval system. Through this system one can access literature (in the form of abstracts), sequences and structures. Entrez is an excellent system for obtaining comprehensive information on a given biological question. The taxonomy browser provides information on taxonomic classification of various species.

The taxonomy database has information on over 79, 000 organisms. Locus link carries information on the official gene names and other descriptive information about genes. Additionally, through Locus link one can access information on homologous genes. For example, it is very convenient to obtain information on the mouse homologue of a given human gene. Homologues from other organisms are also available.

BLAST family of search tools

Among the BLAST (Basic Local Alignment Search Tool) family of similarity search programs are several tools to analyze sequence information. These tools are designed to answer the question "Which sequences in the database are similar (or homologous) to my sequence?" The theory on which BLAST systems were developed is somewhat complex and is out of the scope of discussion here. The principles involved are-

(a) A given sequence is compared with sequences in the database using substitution matrices that specify scores to either 'reward' a match or 'penalize' a mismatch.

(b) Top scoring matches are ranked according to set criteria that serve to distinguish between a similarity due to ancestral relationship or due to random chance. In most analysis these criteria are not changed. However, if the user wishes, criteria can be changed.

(c) True matches are further examined thoroughly with other details accessible through Entrez and other tools available at NCBI.

Note: Two sequences being similar does not mean that they are homologous. Homology is defined as similarity due to common ancestry. Two sequences each from species A and species B are said to be homologous if they have descended from a common ancestor to species A and species B. Duplicated genes within a genome also may have similarities but these are referred to as **'paralogs'**. Homologues will have the same function whereas paralogs may differ in functions.

Resources for gene level sequences

Among the resources for gene level sequences are several tools such as the UniGene, HomoloGene, RefSeq and others. We mentioned about the ESTs in the previous section. The method described therein produces many redundant ESTs because several cDNA clones represent the same gene. To manage the redundancy in EST data, UniGene database was created. The objective is to group ESTs into sets called clusters that belong to 'one' gene (Uni meaning one).

Homologene is a database of orthologs and homologs for several organisms like human, mouse, rat, zebrafish and cow genes represented in UniGene and Locus Link. It is easy to infer homologous relations using this database. RefSeq is a curated database of mRNAs and proteins of organisms like human, mouse and rat. The data provided in RefSeq has been used in many cases such as designing gene chips and describing the sequence features of the human genome.

There are many other resources provided by the NCBI. Discussing all of these is not possible within the space limitations. Now, we mention a few other databases of importance to bioinformatics work (see **Table 6** below)

Table 6. Examples of other useful databases for Bioinformatics.

Database	**Information available**
EMBL(European Molecular Biology Laboratory)	Nucleotide sequence
UniProtKB	Annotated protein sequence
PDB (Protein Database)	Three dimensional structure of proteins
Ribosomal RNA database	rRNA subunit sequences
PALI database	Phylogenetic analysis and alignment of proteins

Curator: A curator is one who reviews and checks newly submitted data ensuring all mandatory information has been provided, that biological features are adequately described and that the conceptual translations of any coding regions obey known translation rules. This process is called curation.

Analysis using Bioinformatics tools

Many kinds of analysis can be made using various bioinformatics tools. These include:

Processing raw information: The experimentally determined sequence (raw information) is processed using bioinformatics tools into genes, the proteins encoded and their function, the regulatory sequences, and inferring phylogenetic relationships.

Genes: Gene prediction can be done by using computer programs like GeneMark for bacterial genomes and GENSCAN for eukaryotes.

Proteins: Protein sequences can be inferred from the predicted genes by using simple computer programs.

Regulatory sequences: Regulatory sequences can also be identified and analysed by using bioinformatics tools.

Inferring phylogenetic relationships: Information regarding the relationships between organisms can be obtained by aligning multiple sequences, calculating evolutionary distance and constructing phylogenetic trees.

Making a Discovery: Using the bioinformatics tools and databases, the functions of unknown genes can be predicted.

Review Questions

1. Why was it necessary to create Bioinformatics database?
2. List the important databases used in routine bioinformatics.
3. What are the IUPAC codes for (i) 'G' or 'C', (ii) A or T, (iii) A or C, (iv) C or T, (v) A or G
4. What are the conventions adopted by the Database personnel to store nucleic acid data and protein sequence data with regard to the direction of the sequence? What is the basis of the convention?
5. What are the single letter IUPAC codes for alanine, glycine, tryptophan, tyrosine, serine, methionine?
6. What are the different types of molecules on which sequence data is obtained and deposited in the database?
7. Name some of the database retrieval tools. What is their purpose?
8. Suggest one possible way for going about analyzing a given sequence using bioinformatics.
9. Using microarrays one can identify the genes expressed differently in normal vs cancer cell types. Explain.
10. What is random shotgun sequencing? What are the difficulties with assembling sequences with repeats?
11. What were the surprises revealed from genome sequencing? What underlies the accrual of complexity in humans even though the number of genes are low?
12. How genes are linked to diseases? Explain with 2 examples.
13. What is BLAST? Describe the principles that underlie BLAST search.
14. What is proteomics? How we can benefit from proteomics?

References

1. Discovering Genomics, Proteomics & Bioinformatics (2002). A Malcolm Compbell and Laurie J. Heyer, published by Benjamin Cummings.

2. Proteomics-from protein sequence to function (2002). S.R. Pennington and M.J. Dunn BIOS Scientific Publishers Limited.

3. Bioinformatics A practical Guide to the Analysis of Genes and Proteins (2002). A.D. Baxeuarus and B.F Francis Ouellehe. A John Wiley & Sons, Inc.