Post-Graduate Degree Programme (CBCS)

in

ZOOLOGY

SEMESTER-III

SOFT CORE THEORY PAPER

HUMAN MOLECULAR GENETICS

ZDSE(MN)T 303

SELF LEARNING MATERIAL

DIRECTORATE OFOPEN AND DISTANCE LEARNING

UNIVERSITY OF KALYANI

KALYANI, NADIA,

W.B., INDIA

Content Writer:

Dr.Subhabrata Ghosh, Assistant Professor of Zoology, Directorate of Open and Distance Learning, University of Kalyani.

Acknowledgements:

The author thankfully acknowledges all the faculty members of Department of Zoology, University of Kalyani for their academic contribution and valuable suggestions regarding the preparation of Self Learning Material.

MAY 2023

Directorate of Open and Distance Learning, University of Kalyani.

Published by the Directorate of Open and Distance Learning,
University of Kalyani, Kalyani-741235, West Bengal.

Director's Message

Satisfying the varied needs of distance learners, overcoming the obstacle of distance and reaching the unreached students are the threefold functions catered by Open and Distance Learning (ODL) systems. The onus lies on writers, editors, production professionals and other personnel involved in the process to overcome the challenges inherent to curriculum design and production of relevant Self Learning Materials (SLMs). At the University of Kalyani a dedicated team under the able guidance of the Hon'ble Vice-Chancellor has invested its best efforts, professionally and in keeping with the demands of Post Graduate CBCS Programmes in Distance Mode to devise a self-sufficient curriculum for each course offered by the Directorate of Open and Distance Learning (DODL), University of Kalyani.

Development of printed SLMs for students admitted to the DODL within a limited time to cater to the academic requirements of the Course as per standards set by Distance Education Bureau of the University Grants Commission, New Delhi, India under Open and Distance Mode UGC Regulations, 2017 had been our endeavour. We are happy to have achieved our goal.

Utmost care and precision have been ensured in the development of the SLMs, making them useful to the learners, besides avoiding errors as far as practicable. Further suggestions from the stakeholders in this would be welcome.

During the production-process of the SLMs, the team continuously received positive stimulations and feedback from Professor (Dr.) Manas Kumar Sanyal, Hon'ble Vice- Chancellor, University of Kalyani, who kindly accorded directions, encouragements and suggestions, offered constructive criticism to develop it within proper requirements. We gracefully, acknowledge his inspiration and guidance.

Sincere gratitude is due to the respective chairpersons as well as each and every member of PGBOS (DODL), University of Kalyani. Heartfelt thank is also due to the Course Writers-faculty members at the DODL, subject-experts serving at University Post Graduate departments and also to the authors and academicians whose academic contributions have enriched the SLMs. We humbly acknowledge their valuable academic contributions. I would especially like to convey gratitude to all other University dignitaries and personnel involved either at the conceptual or operational level of the DODL of University of Kalyani.

Their persistent and co-ordinated efforts have resulted in the compilation of comprehensive, learner-friendly, flexible texts that meet the curriculum requirements of the Post Graduate Programme through Distance Mode.

All the Self Learning Materials are self writing and collected from e-book, journals and websites.

Director
Directorate of Open and Distance Learning
University of Kalyani

ZDSE(MN)T303: HUMAN MOLECULAR GENETICS

Paper	Unit	CONTENT	Credit	Page No
ZDSE(MN)T303 Human Molecular Genetics	I	Human population genetics and evolution: Basic attributes and polymorphic structures in human protein coding genes. Mitochondrial DNA polymorphism. Y-chromosome polymorphism.	1	
	II	Single nucleotide polymorphism (SNP), Regulatory sequence evolution and transposon origin of functional sequences, Basic concept in molecular phylogenetics.		
	III	Genetics in forensic science: Protein comparisons, DNA comparisons, RFLPs, genetic fingerprinting, VNTRs, Genetic profiles.		
	IV	Sociobiology, Altruism, Kin selection and inclusive fitness, Haplodiploidy, Imprinting phenomena.		
	V	Molecular and biochemical basis of genetic diseases: Autosomal (cystic fibrosis), X-linked (haemophilia A), Metabolic disorders (phenylketonuria).		
	VI	Human Genome: Human genome project and the age of genomics, Structure of Human Genome, Concepts and application of Bioinformatics.		

UNIT-I

Human population genetics and evolution: Basic attributes and polymorphic structures in human protein coding genes. Mitochondrial DNA polymorphism and Y-chromosome polymorphism

Objective: In this unit you will know about polymorphism of protein coding genes, Mitochondrial DNA polymorphism and Y-chromosome polymorphism.

Definition of Genetic Polymorphism:

Genetic polymorphism refers to the regular occurrence of several phenotypes in the genetic population. The term genetic polymorphism was coined by Ford in 1940. It has been reported that two third of the loci in a population exhibit polymorphism. The genetic polymorphism is usually maintained due to superiority of heterozygotes over both the homozygotes.

When polymorphism is maintained as a result of heterozygote advantage, it is known as balanced polymorphism. Polymorphism can be detected on the basis of morphological, biochemical and molecular traits or markers.

Genetic polymorphism increases the buffering capacity of a population by providing increased diversity of genotypes in a population. Genetic polymorphism broadens the genetic base of a population and thus enhances the adaptability of the population.

Types of Genetic Polymorphism:

There are six types of genetic polymorphism which are as follows:

i. Balanced Polymorphism:

The genetic polymorphism which is maintained due to superiority of heterozygote over both the homozygotes is referred to as balanced polymorphism. This leads to regular occurrence of several phenotypes in a population. The term balanced polymorphism was first used by Ford in 1940.

Main features of balanced polymorphism are given below:

(a) The polymorphism is maintained due to heterozygote advantage.

(b) This was first reported by Ford in 1940.

(c) This is the most common type of polymorphism observed in plant breeding populations.

ii. Transient Polymorphism:

A genetic polymorphism that is limited to a particular 'period is called transient polymorphism. Sometimes one allele undergoes replacement by a superior allele. The genetic polymorphism during such period is known as transient polymorphism. It is not a regular phenomenon like balanced polymorphism.

Thus there are three main features of transient polymorphism:

(a) It is for a limited period,

(b) It is not a regular feature, and

(c) It was also reported by Ford in 1940.

iii. Neutral Polymorphism:

It refers to the genetic polymorphism that is dependent on a gene action which is almost neutral in its effect on the survival of the genotype in which it is contained. In other words, the effect on the career genotype is neutral. It results due to neutral mutations.

The main features of neutral polymorphism are given below:

(a) It was coined by Ford in 1940.

(b) It results due to neutral mutation.

(c) Neutral mutations take long time in contributing to polymorphism.

iv. Regional Polymorphism:

This refers to occurrence of two or more phenotypes in a population in different regions of the habitat. It results due to adaptation of different individuals in different environment.

Main features of regional polymorphism are given below:

(a) It results due to adaptive variation of alleles.

(b) It is not related to superiority of heterozygotes.

(c) It is also known as geographical polymorphism.

v. Unisexual Polymorphism:

If refers to the genetic polymorphism that is confined to one sex only. It results due to sex limited manifestation of genes. However, such gene can recombine in both sexes.

vi. Cryptic Polymorphism:

It refers to genetic polymorphism in which the genetically different alleles cannot be identified on the basis of their phenotype. It may include chromosomal polymorphism.

Causes of Genetic Polymorphism:

The possible causes of genetic polymorphism include:

i. Heterozygote Advantage:

The natural selection usually favours heterozygotes than homozygotes because heterozygotes are more adaptable than homozygotes. In other words, heterozygotes have more buffering capacity to environmental changes than homozygotes. The heterozygotes maintain genetic polymorphism in a population.

ii. Frequency Dependent Selection:

The frequency dependent selection also leads to maintenance of polymorphism in a population. Generally selection favours those alleles that have low frequency but produce rare phenotype. The selection goes against the alleles that have high frequency. This type of frequency dependent selection maintains balanced polymorphism in a population.

iii. Heterogeneous Environment:

The environment differs from region to region and season to season. The balanced polymorphism is maintained when one allele is advantageous in one environment and another in different environment. In such situation stable polymorphism can be maintained even without heterozygote advantage.

iv. Transition:

In the evolutionary process, sometimes one allele is replaced by another which is more advantageous for adaptation. This may lead to polymorphism in a population. However, such polymorphism is for a limited period and hence is called as transitional polymorphism.

v. Neutral Mutation:

In a population, mutations do arise. However, the majority of mutants are harmful and deleterious. Such mutants are lost only few mutants will survive and replace the original allele. The changes in gene frequency depend on chance. Thus spread of a mutant through the population is erratic.

The frequency of a mutant is sometimes increasing and sometimes decreasing. Only those mutants that have selective advantage will survive and contribute to polymorphism in a population. The surviving few mutants take a very long time to spread in the population but contribute to the polymorphism.

Methods of Detecting Genetic Polymorphism:

In a plant breeding population, the genetic polymorphism can be detected in three main ways, viz.:

(i) On the basis of phenotype

(ii) Biochemical markers and

(iii) Molecular markers.

There are briefly discussed below:

i. On the basis of Phenotype:

The best way of detection of genetic polymorphism is the average heterozygosity at various loci. The higher the heterozygosity, the higher the polymorphism will be. The regular occurrence of several phenotypes in a population is an indication of genetic polymorphism. The polymorphism can be detected on the basis of plant characters such as shape, colour, surface, size of various plant characters. The polymorphism is observed for both oligogenic and polygenic traits.

ii. Biochemical Markers:

Sometimes it is difficult or impossible to identify the polymorphic alleles by visual observations. In such situation the best way of detecting the polymorphic alleles is the isozyme studies or gel electrophoretic studies. Sometimes mutations give rise to protein polymorphism and the variant forms of protein differ only at few amino acid sites. This type of polymorphism can be easily detected by gel electrophoretic studies which throw light on amino acid banding pattern.

iii. Molecular Markers:

Sometimes polymorphic differences are at molecular or DNA level. In other words, the differences are in nucleotide sequences in the DNA. These can be observed by restriction fragment length polymorphisms (RFLPs), amplified fragment length polymorphisms (AFLPs), random amplified polymorphic DNA (RAPD), single sequence repeat (SSR) etc. The molecular or DNA markers are very accurate in detecting the level of polymorphism in a population.

Theories of Genetic Polymorphism:

Two theories have been put forth to explain the wide spread existence of polymorphic variation. These are selectionist theory and neutralist theory.

A brief account of these theories is presented below:

i. Selection Theory:

This theory states that polymorphism is balanced or stable and the stable equilibrium is maintained by selective forces. For example, the balanced polymorphism is the selection in favour of heterozygotes.

ii. Neutral Mutation Theory:

This theory was proposed by Kimura (1983) and further elaborated by Crow (1986). According to this theory some polymorphisms are due to presence of mutant alleles that are nearly neutral with regard to fitness. Such alleles were mutated in distant past and are still present in the population contributing to polymorphism. This theory is widely accepted.

Advantages of Genetic Polymorphism:

There are several advantages of genetic polymorphism which are briefly presented below:

i. Genetic Diversity:

Polymorphic population has greater genetic diversity than pure lines and inbred lines. The genetic diversity avoids danger of uniformity and provides protection from biotic and abiotic stresses to the population.

ii. Broad Genetic Base:

Polymorphic population has broad genetic base due to presence of several phenotypes. Such population has greater buffering capacity to environmental changes.

iii. Adaptation:

Genetic polymorphism enhances the adaptive value of a population by providing increased diversity of genotypes in a population. It also enhances adaptability of a population, because heterozygotes are more adaptable than homozygotes. Genetic polymorphism gives rise to variation of quantitative characters.

Disadvantages of Genetic Polymorphism:

There are some demerits or disadvantages of genetic polymorphism which are briefly discussed below:

i. Difficult to get Purelines:

It is difficult to get purelines from a polymorphic population. Inbreeding does not have much effect in polymorphic population. It is difficult to control the number of loci that have to be kept in polymorphic state.

ii. Less Uniform:

The polymorphic populations are less uniform due to presence of genetic diversity. The produce of such population is also less uniform and less attractive.

iii. Low Yield:

The yield of polymorphic population is poorer than the best genotype present in the polymorphic population.

Modern Concept of Gene:

A gene can be described as a polynucleotide chain, which is a segment of DNA. It is a functional unit controlling a particular trait such as eye colour.

Beadle and Tatum concluded by various experiments that gene is a segment of DNA that codes for one enzyme. They proposed one gene-one enzyme hypothesis. But as some genes code for proteins that are not enzymes, the definition of gene was changed to one gene-one protein hypothesis.

Protein Hypothesis:

The concept of gene has undergone further changes as the new facts came to light. Since proteins are polypeptide chains of amino acids translated by mRNA, gene was defined as one gene-one polypeptide relationship.

Some proteins have two or more different kinds of polypeptide chains, each with a different amino acid sequence. They are products of different genes. For example, haemoglobin has two kinds of chains a andβ chains, which differ in amino acid sequence and length. They are encoded by different genes. Thus, gene is defined as one gene-one polypeptide relationship.

Structural and Regulatory Genes:

Even the one gene-one polypeptide definition is not complete as it does not include gene which codes for rRNA and tRNA. Only mRNA is translated into proteins. Therefore genes which code for polypeptides and RNAs are called structural genes.

In addition to structural genes, DNA also contains some sequences that have only regulatory function. These regulatory genes constitute signals, which "turn on" and "turn off" the transcription of structural genes and perform various other regulatory functions. In this way the definition of gene includes structural genes as well as regulatory genes. Benzer coined terms for the gene, they are Cistron which is the unit of function, Recon which is the unit of recombination and Muton which is the unit of mutation.

Molecular Definition of a Gene:

Gene is defined as the entire nucleic acid sequence that is necessary for the synthesis of a functional gene product, which may be polypeptide or any type of RNA. In addition to structural genes (coding genes) it also includes all the control sequences and non-coding introns. Most prokaryotic genes transcribe polycistronic mRNA and most eukaryotic genes transcribe monocistronic mRNA.

Number of Genes on a Single Chromosome:

Total number of genes on a single chromosome is different in different organisms. Bacteriophage virus R17 consists of only three genes, SV40 consists of 5-10 genes. E. coli bacteria have more than 3000 genes on single 1 mm long chromosome.

Size of a Gene:

In E. coli there are more than four million pairs of nucleotides (4638858 base pairs). It has been estimated that there are about 3000 genes in E. coli.

The minimum size of a gene that encodes a protein can be directly estimated, Each amino acid of a polypeptide chain is encoded by a sequence of three consecutive

nucleotides in a single strand of DNA. Therefore by measuring the size of the polypeptide chain, the size of a gene can be directly measured.

The average polypeptide chain has about 450 amino acids, which are encoded by 1350 nucleotides. Therefore, in E. coli the number of genes will be around 3000 (4000000/1350 = 3000). Human genome contains about 30000 genes, (Source : International Human genome sequencing consortium led in the United States by National Human Genome Research Institute (NHGRI) have estimated the number of human protein coding genes to be less than 30000. Simple round worm C. elegans has about 20000 genes).A single copy of chromosome is composed of more than 3 billion base pairs. Coding regions of these genes take up only 3% of the genome.

Fine Structure of a Gene:

A gene is present only in one strand of DNA, which is a double stranded helix. A gene consists of several different regions. The main region is the coding sequence which carries information regarding amino acid sequence of polypeptides. The region on the left side of coding sequence (upstream or minus region) and on the right side (downstream or plus region) consists of fairly fixed regulatory sequences.

Regulatory sequences consist of promoters which are different in prokaryotes and eukaryotes.

Types of Genes:

1. Simple Genes:

Simple genes have a coding sequence of bases in one DNA strand. Upstream the coding region, the promoter is present. Downstream, the termination region is present.

2. Split Genes:

In most of eukaryotes, many non-coding sequences are present between coding sequences. The coding sequences of DNA of the genes are called exons. In between exons are present non-coding sequences called introns. Exons alternate with introns. Normally introns do not possess any genetic information and are not translated. Such genes are called split genes or interrupted genes.

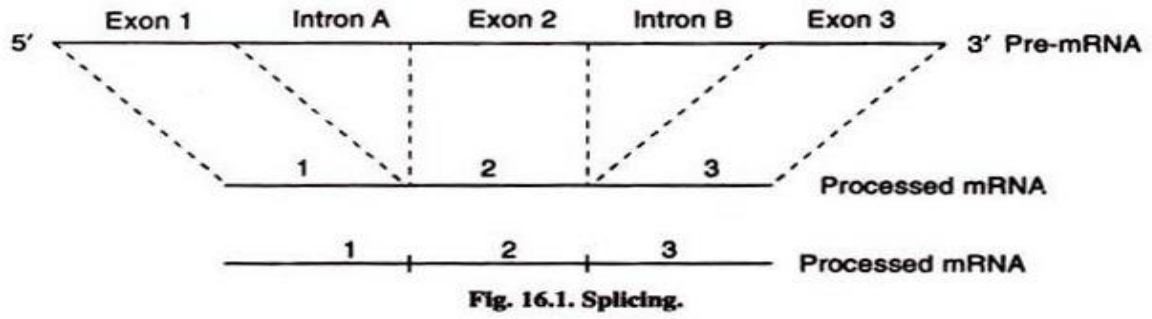


Fig. 16.1. Splicing.

The mRNA transcribed from this DNA is called precursor mRNA (pre-mRNA) and contains exons as well as introns. The introns are removed by excision and discarded. This process is known as splicing. The remaining segments or exons are joined together to form the mature mRNA which takes part in translation. The mature mRNA is much smaller than the pre-mRNA for example α-globin has two introns, ovalbumin has seven introns and α-collagen has 52 introns.

3. Overlapping Genes:

Most genes consist of DNA sequences that code for one protein. But there are some sequences that code for more than one protein. Fredrick Sanger discovered this phenomenon in bacteriophage φ x 174. Overlapping genes are common in many viruses. Here the small length of viral DNA is exploited for synthesizing different proteins.

This is achieved in different ways. In some cases, one gene generates two proteins by having different starting points. Similarly, the same gene generates two proteins by terminating the expression at different points. In other cases, a sequence of DNA makes no distinction between exons and introns. This sequence of DNA, which uses only exons for expression, also uses adjoining introns at other times for expression. The differential splicing of a single stretch of mRNA leads to overlapping and therefore different proteins. In this way, multiple proteins can be generated from a single stretch of DNA.

4. Jumping Genes or Transposons:

Earlier it was thought that genes are static and have definite and fixed locus. However, recently it has been discovered that segments of DNA can jump to new locations in the same or different chromosome. First of all it was discovered by Barbara MClintock in Indian maize corn. It has cobs with kernels of different colours. The light coloured kernels were caused by segments of DNA that move into genes coding for pigmented kernels, thereby inactivating pigmented kernels.

These mobile genes are called transposable elements or transposons. They can jump within the genome, thus affecting the gene expression. Transposable elements are components of moderately repetitive class of DNA.

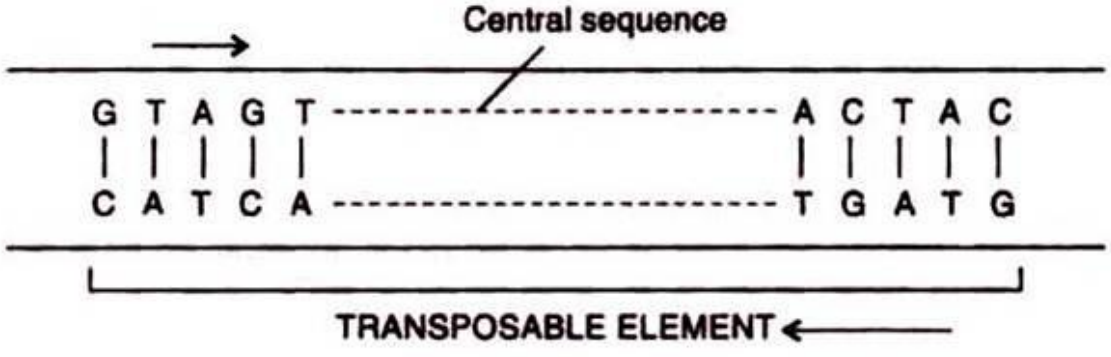


Fig. 16.2.

A transposon has well defined ends. It consists of a long central portion. On either end each transposon has specific sequence of bases which are inverted repeats or palindromes on opposite strands. These terminal repeats help in identifying transposons. The site where a transposon is inserted is called target site or recipient site.Transposable elements can lead to change in the expression of genes. They can also cause mutations. In bacteria, they are present on plasmids.

5. Variable Genes:

Certain polypeptides are coded not by one gene but they are coded by more than one gene present on the same or different chromosomes.

Open Reading Frame:

A gene is a segment of genome which is transcribed into RNA. If the RNA is a transcript of a protein coding gene then it is called messenger RNA or mRNA. This is translated into protein. If the RNA is non-coding as ribosomal RNA (rRNA) or transfer RNA (tRNA) it is not translated.

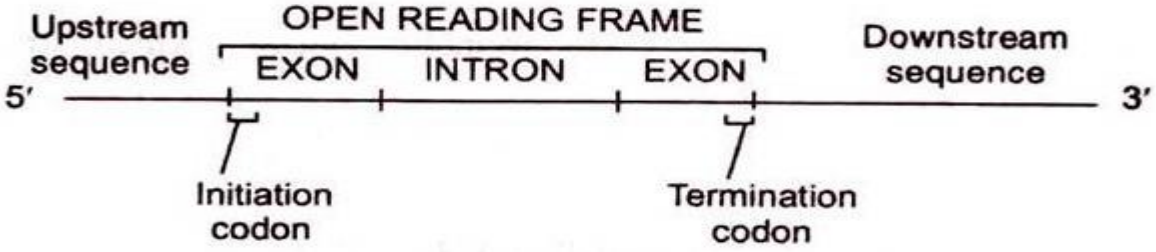


Fig. 16.3.

The part of the protein coding gene which is translated into protein is called open reading frame. It has triplet nucleotide codons. Open reading frame starts with an initiation codon and ends with a termination codon. The region of DNA before a gene is called up-ream region denoted with a minus (-) sign while region after the gene is called downstream denoted with a plus (+) sign. Many genes are split between exons and introns. The introns are removed by splicing to produce a functional RNA before translation.

6. Pseudogenes:

There are some DNA sequences, especially in eukaryotes, which are non-functional or defective copies of normal genes. These sequences do not have any function. Such DNA sequences or genes are known as pseudogenes. Pseudogenes have been reported in humans, mouse and Drosophila.

The main features of pseudogenes are given below:

1. Pseudogenes are non-functional or defective copies of some normal genes. These genes are found in large numbers.

2. These genes being defective cannot be translated.

3. These genes do not code for protein synthesis, means they do not have any significance.

4. The well-known examples of pseudogenes are alpha and beta globin pseudogenes of mouse.

Classification of Genes:

Genes can be classified in various ways. The classification of genes is generally done on the basis of:

(1) Dominance.

(2) Interaction,

(3) Character controlled,

(4) Effect on survival,

(5) Location,

(6) Movement,

(7) Nucleotide sequence,

(8) Sex linkage,

(9) Operon model, and

(10) Role in mutation.

A brief classification of genes on the basis of above criteria is presented in Table 13.4.

TABLE 13.4. Classification and brief description of genes

Classification of genes	*A brief description*
1. Based on Dominance	
Dominant genes	Genes that express in the F_1.
Recessive genes	Genes whose effect is suppressed in F_1.
2. Based on Interaction	
Epistatic gene	A gene that has masking effect on the other gene controlling the same trait.
Hypostatic gene	A gene whose expression is masked by another gene governing the same trait.
3. Based on Character Controlled	
Major gene	A gene that governs qualitative trait. Such genes have distinct phenotypic effects.
Minor gene	A gene which is involved in the expression of quantitative trait. Effect of such genes cannot be easily detected.
4. Based on Effect on Survival	
Lethal gene	A gene which leads to death of its carrier when in homozygous condition. It may be dominant or recessive.
Semilethal gene	A gene that causes mortality of more than 50% of its carriers.
Sub-vital gene	A gene that causes mortality of less than 50% of its carriers.
Vital gene	A gene that does not have lethal effect on its carriers.
5. Based on Location	
Nuclear genes	Genes that are found in nuclear genome in the chromosomes.
Plasma genes	Genes that are found in the cytoplasm in mitochondria and chloroplasts. Also called cytoplasmic or extranuclear genes.
6. Based on Position	
Normal genes	Genes that have a fixed position on the chromosomes. Most of the genes belong to this category.
Jumping genes	Genes which keep on changing their position on the chromosome of a genome. Such genes have been reported in maize.
7. Based on Nucleotide Sequence	
Normal genes	Genes having continuous sequence of nucleotides which code for a single polypeptide chain.
Split gene	A gene having discontinuous sequence of nucleotides. Such genes have been reported in some eukaryotes. The intervening sequences do not code for amino acids.
Pseudo genes	Genes having defective nucleotides which are non-functional. These genes are defective copies of some normal genes.
8. Based on Sex Linkage	
Sex linked genes	Genes which are located on sex or X-chromosomes.
Sex limited genes	Genes which express in one sex only.

Polymorphism of Mitochondrial DNA:

Mitochondrial DNA is a double stranded circular molecule, which is inherited from the mother in all multi-cellular organisms, though some recent evidence suggests that in rare instances mitochondria may also be inherited via a paternal route. Typically, a sperm carries mitochondria in its tail as an energy source for its long journey to the egg. When the sperm attaches to the egg during fertilization, the tail falls off. Consequently, the only mitochondria the new organism usually gets are from the egg its mother provided. There are about 2 to 10 transcripts of the mt-DNA in each mitochondrion. Compared to chromosomes, it is relatively smaller, and contains the genes in a limited number.

The size of mitochondrial genomes varies greatly among different organisms, with the largest found among plants, including that of the plant Arabidopsis, with a genome of 200 kbp in size and 57 protein-encoding genes. The smallest mtDNA genomes include that of the protist Plasmodium falciparum, which has a genome of only 6 kbp and just 2 protein- encoding genomes. Humans and other animals have a mitochondrial genome size of 17 kbp and 13 protein genes.

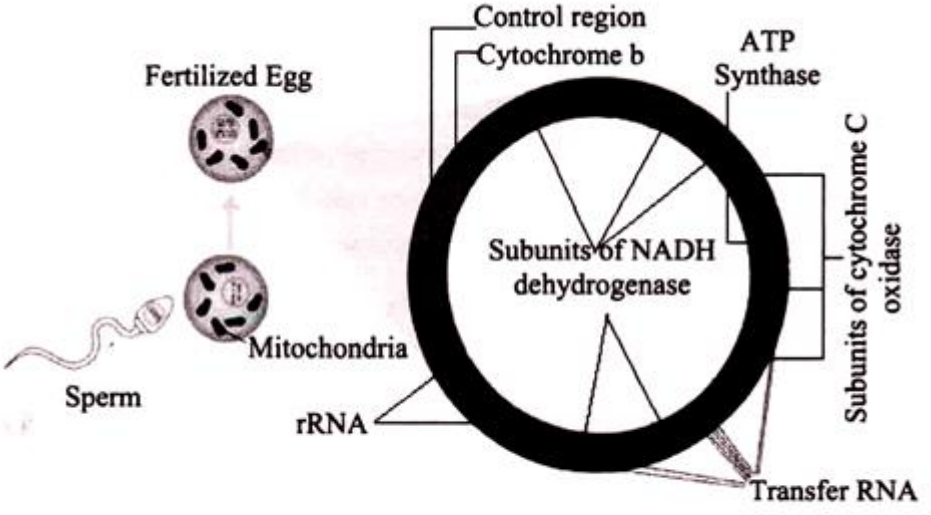


Figure 4.56: Mitochondrial DNA

Mitochondrial DNA consists of 5-10 rings of DNA and appears to carry 16,569 base pairs with 37 genes (13 proteins, 22 t-RNAs and two r-RNA) which are concerned with the production of proteins involved in respiration. Out of the 37 genes, 13 are responsible for making enzymes, involved in oxidative phosphorylation, a process that uses oxygen and sugar to produce adenosine tri-phosphate (Fig. 4.56). The other 14 genes are responsible for making molecules, called transfer RNA (t-RNA) and ribosomal RNA (r-RNA). In some metazoans, there are about 100 – 10,000 separate copies of mt-DNA present in each cell.

Unlike nuclear DNA, mitochondrial DNA doesn't get shuffled every generation, so it is presumed to change at a slower rate, which is useful for the study of human evolution. Mitochondrial DNA is also used in forensic science as a tool for identifying corpses or body parts and has been implicated in a number of genetic diseases, such as Alzheimer's disease and diabetes. Changes in mt-DNA can cause maternally inherited diseases, which leads to faster aging process and genetic disorders.

Mitochondria convert the potential energy of food molecules into ATP by the Krebs cycle, electron transport and oxidative phosphorylation in presence of oxygen. The energy from food molecules (e.g., glucose) is used to produce NADH and $FADH_2$ molecules, via glycolysis and the Krebs cycle. The protein complexes in the inner membrane (NADH dehydrogenase, cytochrome c reductase, cytochrome c oxidase) use the released energy to pump protons (FT) against a gradient.

Mitochondrial DNA:

Each human cell contains hundreds of mitochondria each containing multiple copies of mitochondrial DNAs (mtDNA). Mitochondria generate cellular energy through the process of oxidative phosphorylation. As a by-product they produce most of the endogenous toxic reactive oxygen species Mitochondria are also the central regulators of apoptosis or programmed cell death.

These interrelated functional systems involve activities of about 1000 genes distributed in the nuclear genome and the mitochondrial genome. Due to their dependence on the nuclear genome, mitochondria are considered as semi-autonomous.This has been shown by experiments in which mitochondria and mtDNA could be transferred from one cell to another. The donor cell was enucleated and its mitochondria-containing cytoplast fused with a recipient cell (technique of cybrid transfer).

The genomes of mitochondria show wide variation particularly among plants and protists. Most mitochondrial DNAs (mtDNA) consist of a closed circular double stranded supercoiled DNA molecules located in multiple nucleoid regions (similar to those in bacterial cells); some protists however, have varying lengths or multiple circular molecules of DNA as in the trypanosomes. mtDNA in the protist *Amoebidium parasiticum* consists of several distinct types of linear molecules with terminal and sub-terminal repeats. Although most mtDNAs are in the size range of 15 to 60 kb, mtDNA in malarial parasite (*Plasmodium* spp) is only 6 kb long, while that of rice (Oryza sativa) is 490 kb, and cucurbits 2 Mb. There are about 40 to 50 coding genes in mitochondrial DNA, Plasmodium being an exception with 5 coding genes.

The large size of mitochondrial genome in plants is due to noncoding inter-genic regions and their content of tandem repeats. Introns are present in many mtDNAs, and in some unusual cases, the genes are split into as many as 8 regions that are dispersed in the genome, and located on both strands of the DNA. Transcription takes place separately in portions of the split genes producing discrete pieces of RNA that are held together by base pairing of complementary sequences. The mtDNA contains information for a number of mitochondrial compounds such as tRNAs, rRNA, and some of the polypeptide subunits of the proteins cytochrome oxidase, NADH- dehydrogenase and ATPase. Most of the other proteins found in mitochondria are encoded by the nuclear genome and transported into mitochondria.These include DNA polymerase and other proteins for mtDNA replication, RNA polymerase and other proteins for transcription, ribosomal proteins for ribosome assembly, protein factors for translation, and the aminoacyl-tRNA synthetases.The mitochondrial oxidative phosphorylation complexes are composed of multiple polypeptides, mostly encoded by the nuclear DNA (nDNA). However, 13 polypeptides are encoded by mtDNA. The mtDNA also codes for 12S and 16S rRNAs and 22 tRNAs required for mitochondrial protein synthesis. The mtDNA also contains a control region consisting of approximately 1000 base pairs constituting the promoter region and the origin of replication.

The mRNAs synthesised within the mitochondria remain in the organelle and are translated by mitochondrial ribosomes that are assembled within mitochondria. Mitochondrial ribosomes have two subunits. Mitochondria in human cells have 60S ribosomes consisting of a 45S and a 35S subunit.There are only two rRNAs in mitochondrial ribosomes of most organisms, that is, 16S rRNA in large subunit and 12S rRNA in small subunit of most animal ribosomes. There is usually one gene for each rRNA in a mitochondrial genome. The proteins in mitochondrial ribosomes are encoded by the nuclear genome and transported into mitochondria from the cytoplasm.

Mitochondrial ribosomes are sensitive to most of the inhibitors of bacterial ribosome function such as streptomycin, neomycin and chloramphenicol. For protein synthesis, mitochondria of most organisms use a genetic code that shows differences from the universal genetic code. Only plant mitochondria use the universal nuclear genetic code. Transcription of mammalian mtDNA is unusual in that each strand is transcribed into a single RNA molecule that is then cut into smaller pieces. In the large RNA transcripts that are produced, most of the genes encoding the rRNAs and the mRNAs are separated by tRNA gene.

The tRNAs in the transcript are recognised by specific enzymes and are cut out, leaving only the mRNAs and the rRNAs. A poly (A) tail is then added to the 3'end of each mRNA and CCA is added to the 3'end of each tRNA. There are no 5′ caps in mitochondrial mRNAs.Mitochondrial DNA replication is semi-conservative and uses DNA polymerases that are specific to the mitochondria. The mtDNA replicates throughout the cell cycle, independently of nuclear DNA synthesis which takes place in S phase of cell cycle. Observations on mtDNA replication in animal mitochondria in vivo have resulted in a model referred to as the displacement loop (D loop) model as follows (Fig. below).

The two strands of mtDNA in most animals have different densities because the bases are not equally distributed on both strands, called H (heavy) and L (light) strands. The synthesis of a new H strand starts at the replication origin for the H strand and forms a D-loop structure (Fig. below).As the new H strand extends to about halfway around the molecule, initiation of synthesis of a new L strand takes place at a second replication origin. Synthesis continues until both strands are completed. Finally, each circular DNA assumes a supercoiled form.

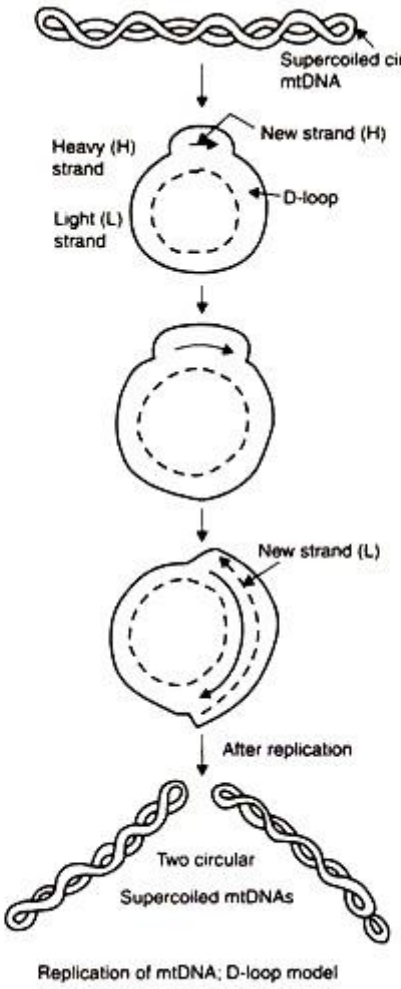


Fig. 17.6 Model for mitochondrial DNA replication by formation of a D-loop structure.

The mtDNA is maternally inherited and has a very high mutation rate. When a new mtDNA mutation occurs in a cell, a mixed intracellular population of mtDNAs is generated, known as heteroplasmy. During replication in a heteroplasmic cell, the mutant and normal molecules are randomly distributed into daughter cells.

When the percentage of mutant mtDNAs increases, the mitochondrial energy producing capacity declines, production of toxic reactive oxygen species increases cells become more prone for apoptosis. The result is mitochondrial dysfunction. Tissues most sensitive to mitochondrial dysfunction are brain, heart, kidney and skeletal muscle.ThemtDNA mutations are associated with a variety of neuromuscular disease symptoms, including various ophthalmological symptoms, muscle degeneration, cardiovascular diseases, diabetes mellitus, renal function and dementias.

The mtDNA diseases can be caused either by base substitutions or rearrangement mutation. Base substitution mutations can either alter protein (missense mutation) or rRNAs and tRNAs (protein synthesis mutations). Rearrangement mutations generally delete at least one tRNA and thus cause protein synthesis defects.Missense mutations are associated with myopathy, optic atrophy, dystonia and Leigh's syndrome. Base substitution mutations in protein synthesizing genes have been associated with a wide spectrum of neuromuscular diseases, and the more severe typically include

mitochondrial myopathy.Mitochondrial diseases are also associated with a number of different nuclear DNA mutations. Mutations in the RNA component of the mitochondrial RNase have been implicated in metaphyseal chondrodysplasia or cartilage hair hypoplasia which is an autosomal recessive disorder resulting from mutation in nuclear chromosome 9 short arm position (9p13).

Y-Chromosome Polymorphism:

The Y chromosome is one of two sex chromosomes (allosomes) in mammals, including humans, and many other animals. The other is the X chromosome. Y is the sex-determining chromosome in many species, since it is the presence or absence of Y that determines the male or female sex of offspring produced in sexual reproduction. In mammals, the Y chromosome contains the gene SRY, which triggers testis development. The DNA in the human Y chromosome is composed of about 59 million base pairs. The Y chromosome is passed only from father to son. With a 30% difference between humans and chimpanzees, the Y chromosome is one of the fastest-evolving parts of the human genome. To date, over 200 Y-linked genes have been identified. All Y-linked genes are expressed and (apart from duplicated genes) hemizygous (present on only one chromosome) except in the cases of aneuploidy such as XYY syndrome or XXYY syndrome.

The following are some of the gene count estimates of human Y chromosome. Because researchers use different approaches to genome annotation their predictions of the number ofgenes on each chromosome varies (for technical details, see gene prediction). Among various projects, the collaborative consensus coding sequence project (CCDS) takes an extremely conservative strategy. So CCDS's gene number prediction represents a lower bound on the total number of human protein-coding genes.In human genetics, a human Y-chromosome DNA haplogroup is a haplogroup defined by mutations in the non-recombining portions of DNA from the Y-chromosome (called Y-DNA). Mutations that are shared by many people are called single-nucleotide polymorphisms (SNPs). Haplogroups are defined through mutations (SNPs).

The human Y-chromosome accumulates roughly two mutations per generation. Y-DNA haplogroups represent major branches of the Y-chromosome phylogenetic tree that share hundreds or even thousands of mutations unique to each haplogroup.The Y-chromosomal most recent common ancestor (Y-MRCA, informally known as Y chromosomal Adam) is the most recent common ancestor (MRCA) from whom all currently living men are descended patrilineally. Y-chromosomal Adam is estimated to have lived roughly 236,000 years ago in Africa. By examining other bottlenecks most Eurasian men are descended from a man who lived 69,000 years ago. Other major bottlenecks occurred about 5,000 years ago and subsequently most Eurasian men can trace their ancestry back to a dozen ancestors who lived 5,000 years ago.

Y-DNA haplogroups are defined by the presence of a series of Y-DNA SNP markers. Subclades are defined by a terminal SNP, the SNP furthest down in the Y-chromosome phylogenetic tree. The Y Chromosome Consortium (YCC) developed a system of naming major Y-DNA haplogroups with the capital letters A through T, with further subclades

named using numbers and lower case letters (YCC longhand nomenclature). YCC shorthand nomenclature names Y DNA haplogroups and their subclades with the first letter of the major Y-DNA haplogroup followed by a dash and the name of the defining terminal SNP. Y-DNA haplogroup nomenclature is changing over time to accommodate the increasing number of SNPs being discovered and tested, and the resulting expansion of the Y-chromosome phylogenetic tree. This change in nomenclature has resulted in inconsistent nomenclature beingused in different sources. This inconsistency, and increasingly cumbersome longhand nomenclature, has prompted a move towards using the simpler shorthand nomenclature.

Recent innovations have included the creation of primers targeting polymorphic regions on the Y-chromosome (Y-STR), which allows resolution of a mixed DNA sample from a male and female or cases in which a differential extraction is not possible. Y-chromosomes are paternally inherited, so Y-STR analysis can help in the identification of paternally related males. Y-STR analysis was performed in the Sally Hemings controversy to determine if Thomas Jefferson had sired a son with one of his slaves. The analysis of the Y-chromosome yields weaker results than autosomal chromosome analysis. The Y male sex-determining chromosome, as it is inherited only by males from their fathers, is almost identical along the patrilineal line. This leads to a less precise analysis than if autosomal chromosomes were testing, because of the random matching that occurs between pairs of chromosomes as zygotes are being made

Probable Questions:

1. What do you mean by polymorphism? Give suitable examples.
2. Define balanced polymorphism. What are the main features of it?
3. Define transient polymorphism. What are the main features of it?
4. Define neutral polymorphism. What are the main features of it?
5. Define regional polymorphism. What are the main features of it?
6. What are the causes of genetic polymorphism?
7. Define neutral mutation theory and selection theory.
8. What are the advantages of genetic polymorphism?
9. What are the disadvantages of genetic polymorphism?
10. Define split genes and overlapping genes.
11. Define pseudogenes and jumping genes.
12. Write a brief note on mitochondrial DNA polymorphism.
13. Write a brief note on Y chromosome DNA polymorphism.

Suggested Readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics . Verma and Agarwal

UNIT-II

Single nucleotide polymorphism (SNP) Regulatory sequence evolution,basic concept in molecular phylogenetics

Objective: In this unit we will discuss about Single nucleotide polymorphism (SNP), Regulatory sequence evolution and transposon origin of functional sequences, Basic concept in molecular phylogenetics.

Single Nucleotide Polymorphism (SNPs):

A Single Nucleotide Polymorphism or SNP (pronounced 'snip') is a small genetic change, or variation, that can occur within a DNA sequence.

The four nucleotide letters A (adenine), C (cytosine), T (thymine), and G (guanine) specify the genetic code.

SNP variation occurs when a single nucleotide, such as an A, replaces one of the other three nucleotide letters – C, G, or T (Fig. 20-1).

By classical definition of polymorphism the frequency of the variation will have to be at least 1% to qualify the nucleotide change as a polymorphism. Those nucleotide changes that occurs less than 1% would be called rare variant.

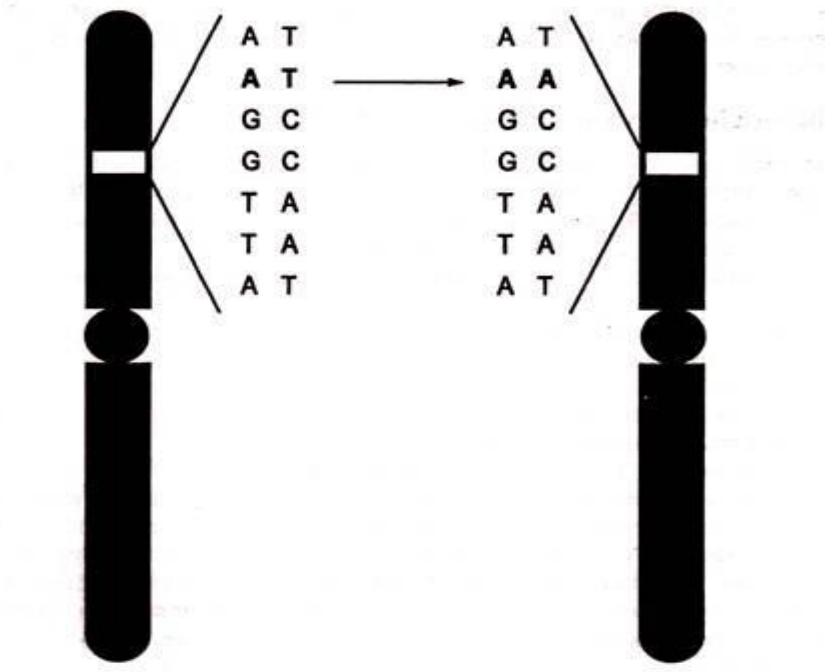


Fig. 20-1 : *A single nucleotide polymorphism (SNP) is shown (A>T) in a stretch of DNA sequence in the chromosome*

Because only about 1.1 to 1.4% of a person's DNA sequences codes for proteins, most SNPs are found outside of coding sequences. SNPs lying outside the coding region normally would not be expected to have any impact on the phenotype of an organism. SNPs found within a coding sequence are of particular interest to researchers as they are more likely to alter the biological function of a protein, although these changes have much less drastic effect than that of mutations.
Due to recent advances in field of gene identification and characterization, there has been a huge flurry of SNP discovery. Finding single nucleotide changes throughout the human genome seems a mammoth job, but, over the last 20 years, researchers have developed a number of techniques that makes it possible.

Each technique uses a similar non-identical method to compare selected regions of a DNA sequence obtained from multiple individuals who share a common trait. In each test, the result shows a difference in the DNA samples when a SNP is detected in one individual in a pool under test.

Distribution of SNPs:

SNPs are not distributed uniformly over the genome. A huge number of SNPs are distributed throughout the non-coding region of the genome. Since these regions are free from selection pressure, these changes are selected neutrally and fixed over time. The distribution patterns of the SNPs are variable even in a single chromosome.

For instance, regions responsible for antigen presentation to the immune system, present on the chromosome 6, shows very high nucleotide variability in contrast to other regions of the same chromosome.

The Origin, Survival and Fixation of SNPs:

The SNP is the main source of variance in the genome and it accounts for 90% of all human polymorphism.

There are Two Types of Nucleotide Base Substitution:

Transition:

Transformation, which accounts for nearly two-thirds of all SNPs, occurs between purines (e.g. A > G) or pyrimidines (e.g. C > T).

Transversion:

Transversion occurs between purines and pyrimidines (e.g. A > C and G > T).

A SNP undergoes series of selection procedures before finally being established.

Its Life can be Roughly Divided into 4 Phases:

1) Appearing by the means of point mutations.

2) Surviving the selection pressure of the nature.

3) Spreading through generations.

4) Establishing itself at least as 1% of all alleles.

The most frequent change in humans is the mutation from CpG to TpG (a transition accounting for approximately 25% of all mutations). This mechanism causes decrease in the number of CG dinucleotide in the genome since many eventually becomes TG, whereas new CpG sites will be created by other less frequent mutations.

Since only 1.1% to 1.4% of the genome codes for proteins SNPs are likely to occur at non-coding sequences more frequently. Even if the SNP occurs at a coding sequence, mostly it might have a subtle and non-deleterious effect on the expressed proteins. Changes accounting for deleterious effects are eventually removed from the genome by natural selection. Hence to attain the status of an SNP, a point mutation should be non-deleterious to be selected (Miller and Kwok, 2001).

Genetic Predisposition:

Most common diseases in humans are not caused by a genetic variation within a single gene, but are influenced by complex interactions among multiple genes as well as environmental and lifestyle factors. Although both environmental and lifestyle factors add up in the phenotype of a disease, it is difficult to measure and evaluate their overall effect on a disease process.

The probability of an individual to develop a disease based on genes and hereditary factors is referred to as genetic predisposition. Genetic factors confer susceptibility or resistance to a disease and determine the severity or progression of disease.Most of the predisposition factors are still unknown. Researchers have found it difficult to develop screening tests for most diseases and disorders. Phenotypic association of certain coding SNPs with a disorder of that specific gene has led to identification of functional aspect of the SNPs.

Single Nucleotide Polymorphism also can be used as a tool for identifying genes, responsible for the disease or, genes imparting a certain phenotype of the disease. By studying stretches of DNA sequence that have been found to harbor a SNP associated with a disease trait, researchers may begin to reveal relevant genes associated with a disease.Understanding the role of genetic factors in disease will also allow researchers to better evaluate the role of non-genetic factors—such as habitat, upbringing, behavior, diet, lifestyle, and physical activity, on the disease.

As genetic factors also affect an individual's response to a drug therapy, SNPs will be useful in helping researchers determine and understand why individuals differ in their abilities to metabolize certain drugs, as well as to determine why an individual may experience an adverse side effect to a particular drug.Therefore, the recent discovery of

SNPs promises to revolutionize not only the process of disease detection, but also the practice of personalized, preventative and curative medicine.

Application of SNP in Pharmacogenomics Studies:

Response rates towards major and common drugs vary overtly among individuals (Table 1). SNPs attribute in a major way towards this phenomenon. Using SNPs to study the genetics of drug response has the potential to help in the creation of personalized medicine as explained in (Fig. 20.2). As mentioned earlier, SNPs may also be associated with the metabolism i.e., absorbance and clearance of therapeutic agents.

Currently, there is no standard genetic screening of drug metabolizing genes to determine how a patient will respond to a particular medication. A treatment proven effective in one patient may be ineffective in others. Some patients may also experience adverse immunological reaction to a particular drug.

Hence pharmaceutical companies limit their production of drugs for which an 'average' patient will respond. As a result a relatively smaller group of patients harboring any putative genetic variation (e.g. a SNP), which renders them unable to metabolize that drug, remains untreated. Many drugs that might benefit that small group of patients never make it to market as those drugs would fetch less profit for the drug Industries.

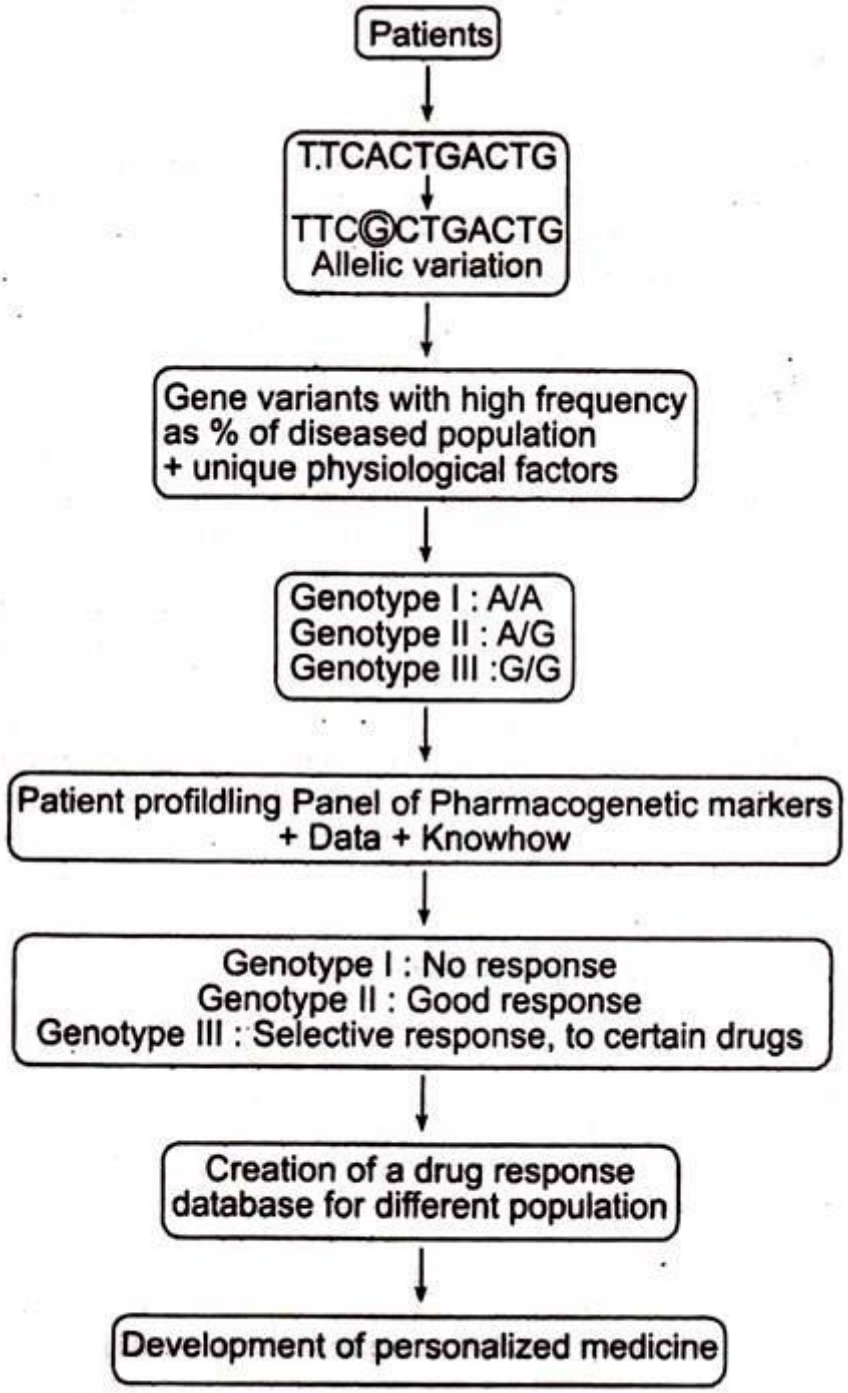


Fig. 20-2 : *A flow diagram showing use of genetic information in determination of efficacy of medicine in individuals based on the SNP profile*

Table 1 : RESPONSE RATES OF PATIENTS TO A MAJOR DRUG FOR A SELECTED GROUP OF THERAPEUTIC AREAS

Therapeutic Area	Efficacy Rate (%)
Alzheimer's	30
Analgesics (Cox-2)	80
Asthma	60
Cardiac arrhythmia	60
Depression (SSRI*)	62
Diabetes	57
Hepatitis C Virus	47
Incontinence	40
Migraine : acute / prophylaxis	52/50
Oncology	25
Osteoporosis	48
Rheumatoid arthritis	50
Schizophrenia	60

*SSRI= Selective Serotonin Reuptake Inhibitors

The data presented in the above Table (taken from the Physicians Desk Reference, 54th edn., Medical Economics Company, 2000) shows the limitation of the efficacy of prescribed drugs to ameliorate the disease among the affected individuals and underscores the importance of exploring for personalized medicine based on the genetic make up of the individuals.

The post-genomic era has revealed association of SNP with certain human diseases either directly or indirectly. For example, genetic studies have shown intricate relationship between:

(a) SNPs in coagulation factor gene F5 and deep-vein thrombosis,

(b) Genetic alteration in the chemokine receptor gene CCR5 and susceptibility to HIV infection and relation between a host of other SNPs and diseases (McCarthy and Hilfiker, 2000).

These associations exemplify the candidate gene approach and so can be beneficial to identify the condition of predisposition to the disease in an individual (Table 2). Similar, associations are also seen among SNP and drug response variations in individuals (Table 3). For example, genetic variants in a drug-metabolizing enzyme (thiopurine methyltransferase ; TPMT) have been linked to adverse drug reactions (Snow & Gibson, 1995); similarly variation at ALOX5 promoter modulates the response to anti-asthma treatment (Drazen et al. 1999). SNPs in Apolipoprotein E (APOE) gene have been associated with response towards cholinesterase inhibitor in Alzheimer's patients.

Direct effects of SNP are also seen in various common diseases. Recently SNPs responsible for increased risk of diabetes have been identified. A common genetic variant due to a SNP is peroxisome-proliferator-activated-receptor (PPAR) gamma gene, present in around 25% of type 2 diabetes patients in the population, is thought to make

individuals more prone to diabetes (Altshuler et al. 2000).Although these variations have a modest effect on individual risk but still affect a major portion of the human population. In contrast to identification of direct involvement of SNPs in disease, identification and use of SNPs in gene responsible for drug metabolism and detoxification is limited.

As drug response legends upon the administration of the drug, 'responders and non-responders' can only identified after they receive the drug. This makes identification of individuals from a population difficult (McCarthy and Hilfiker, 2000). A correlative study of effect of a drug on a large population already known, followed by a SNP screening of suspect genes, followed by statistical interpretation, may turn out to be useful in understanding SNP-phenotype relationship.

Table 2 : EXAMPLES OF PHARMACOLOGICALLY RELEVANT POLYMORPHISMS

Gene	Mutation/Variant	Effect	Reference
Dopamine D5 receptor (*DRD5*)	Asn351 AsP	Approximately 10-fold decrease in dopamine and 3-fold decrease in R(+)–SKF 38393 binding affinities	Cravchik and Gejman, 1999
Dopamine D2 receptor (*DRD2*)	Ser311 Cys, Pro310Ser and Val96Ala	Alteration in binding affinities to *DRD2* and potency of several neuroleptics commonly used in the treatment of psychotic disorders	Cravchik *et al.* 1999
Apolipoprotein E (*ApoE*)	APoE2, E3 and E4 alleles	The ApoE4 allele is a prognostic indicator of poor response to therapy with acetyl-cholinesterase inhibitor in Alzheimer's patients	Farlow *et al.* 1998 ; Poirier *et al.* 1995
Herceptin (*HER2*)	Normal expression/over expression variants	*HER2* over-expression results in a more aggressive, less responsive breast cancer. *HER2* over-expression is linked to sensitivity and/or resistance to hormone therapy and chemotherapeutic regimens	Mitchell and Press, 1999

In future, the most appropriate drug for an individual could be determined and used for treatment by analyzing a patient's SNP profile. The ability to target a drug to those individuals most likely to benefit, referred to as 'personalized medicine' would allow pharmaceutical companies to bring many more drugs to market and allow doctors to prescribe individualized therapies specific to a patient's needs.

The information can be integrated with other resources such as structure of proteins. Using a computational approach, coding SNPs can be plotted on the protein 3D structure and the observed changes can be correlated with the phenotype.

The application of structural data to research on genetic variation is of immense use for studies on the genetic basis of phenotypic variation. Pharmaceutical industry can profit from drug metabolizing gene 'SNP and mutation database' containing information regarding the structure of the polymorphism, percentage in the population bearing the variant genotypes, and its phenotypic effect in response to drug treatment.

Structural and Functional Importance of SNPs:

In addition to the SNPs occurring in the coding sequence of genes, functional importance of SNPs has also been observed in non-coding DNA (e.g. introns) including regulatory (e.g. promoters, enhancers etc). One good example of functional SNP is in a non-coding region is the tau gene. The structure of tau exon 10 splicing regulatory element RNA has been recently deciphered and has been shown to form a stable folded stem-loop structure.

Other examples are:

Intronic SNP affecting Splice sites:

Coding region and intronic mutations in the tau gene cause frontotemporal dementia and Parkinsonism linked to chromosome 17. Intronic mutations and some missense mutations increase splicing in-of exon 10, leading to an increased ratio of four-repeat to three-repeat tau isoforms (Varani et al. 1999).

Table 3 : CLINICALLY RELEVANT GENETIC POLYMORPHISMS THAT INFLUENCE DRUG METABOLISM

Gene	Drug : Therapy	Clinical Response	Reference
		DRUG METABOLIZING ENZYMES	
CYP2C9	Warfarin : anticoagulation	Dosing in patients with R144C allele (reduced catalytic activity) use lower maintenance dose for anti-coagulation therapy	Furuya *et al.* 1995
CYP2D6	Codeine : analgesic	Patients with two inactive alleles do not metabolize codeine to morphine and get no analgesia	Sindrup and Brosen, 1995
Thiopurine methyl-transferase	Thipurines : leukemia, autoimmune disorders	Patients with homozygous G460A and A719G can develop toxic overdose in azathioprine therapy	Snow and Gibson 1995
		DRUG TARGETS	
β-2 Adrenergic receptor	Albuterol : asthma	Patients homozygous for Gly17Arg mutations suffer exacerbation of asthma symptoms with regular use of albuterol	Israel *et al.* 2000
ALOX-5 (5-lipoxygenase	Zileuton : asthma	Patients with two non-expressing alleles of Alox-5 do not respond to 5-lipoxygenase inhibitor	Drazen *et al.* 1999

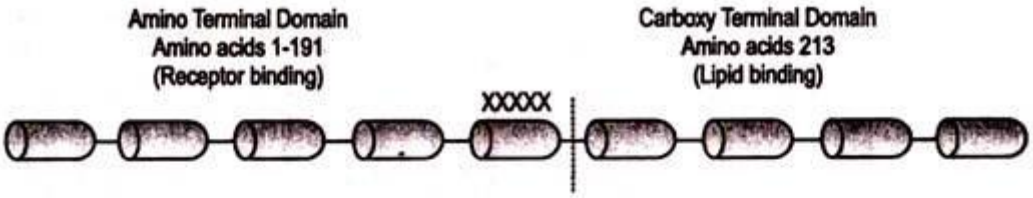
Promoter SNPs Affecting Gene Expression:

SNPs can affect gene expression if they happen to lie on the promoter or any other control sequence of the gene. Transcription factors and the RNA polymerase bind differentially to the promoters based on the sequence context, influencing the expression pattern of the gene. Recent evidences show that SNPs in the promoter region of TNF-α, IL-1β and few other cytokines enhances its expression, hence rendering the individuals more immune to some bacterial and other pathogenic infections (El-Omar et al. 2000).

SNPs in the Coding Regions Affecting the Protein Structure:

Subtle alteration in the DNA could result in drastic alteration in the protein structure as discussed. Structural and functional relationship of apolipoprotein (apo) E in lipoprotein metabolism, heart disease, and neurodegenerative diseases, including Alzheimer's disease has been established. ApoE is a 299 amino acid long protein with two functional domains.

The amino-terminal domain containing the residues 1-191 contains the low density lipoprotein (LDL) receptor- binding-region, and the carboxyl-terminal domain contains the major lipid-binding elements.The three common human isoforms—apoE2, apoE3, and apoE4—differ only at two positions in the protein but have very different metabolic properties and dramatic impacts on disease (Fig. 20.3). ApoE3 (Cys-112, Arg-158) binds normally to the LDL receptor and is associated with normal lipid metabolism, whereas apoE2 (Cys-112, Cys- 158) binds defectively to the LDL receptor and is associated with the genetic disorder type III hyperlipoproteinemia. ApoE4 (Arg-112, Arg-158) binds normally to the LDL receptor but is associated with elevated cholesterol levels and, hence, an increased risk for cardiovascular disease (Morrow et al. 2002). In addition, it has been observed that apoE4 is a major risk factor for Alzheimer's disease.



Isoforms	Amino acid variants		Associated phenotype
E2	Cys 112	Cys 158	Type III hyperlipoproteinemia
E3	Cys 112	Arg 158	Normal lipid metabolism
E4	Arg 112	Arg 158	Elevated cholesterol level

Fig. 20-3 : *Association of apolipoprotein E (ApoE) to various diseases. A structural model demonstrates the helices (cylinders) and random coils (lines) in the ApoE protein and the critical binding residues which binds to low density lipoprotein (LDL) represented by 'X's. E2, E3 and E4 are three isoforms varying at two residues of the ApoE protein*

A number of recent case studies on the effect of SNP on the structure and function of proteins have not only shown the specific structural alteration of the protein in disease, but have also given insights into the regulatory mechanisms of the native protein. Presence of a point mutation (Leu55Pro) in α1-antichymotrypsin, a protease inhibitor of the serpin superfamily, causes its loss of activity (Sunyaev et al. 2001).

The change in the protein due to the point mutation causes obstructive pulmonary disease. Similarly, a point mutation in human apolipoproteinA-1 (ApoA-1) is associated with coronary heart disease (Sunyaev et al. 2001). These studies on the effects of single nucleotide changes on protein structure gives us insights into both the cause of disease and the functions of protein.

Similar studies have been done in mu-opioid receptor, oprm 1. The mu-opioid receptor is the primary site of action for the most commonly used opioids, including morphine, heroin, fentanyl, and methadone. The most prevalent SNP present in about 10% of the population is a nucleotide substitution at position 118 (118 A > G), with predicted amino acid change at a putative N-glycosylation site.Although the variant protein resulting from the 118 A > G SNP did not show altered binding affinities for most opioid peptides and alkaloids tested, the 118 A > G variant receptor bound beta-endorphin an opioid that activates the opioid receptor, binds approximately 3 times more tightly than the most common allelic form of the receptor.

Furthermore, beta-endorphin is approximately 3 times more potent at the 118 A > G variant receptor than at the common allelic form in agonist-induced activation of G protein-coupled potassium channels. These results suggested that 118 A > G SNP in the opioid receptor gene may have implications for normal physiology and vulnerability to develop diverse diseases including the addictive diseases (Bond et al. 1998).

Relation of SNPs with blood pressure (BP) has also been established. Single nucleotide changes in the Angiotensinogen (AGT) gene are observed, which is common in people with high blood pressure. A North American religious genetic isolate, Hutterites, was tested for association between variation in systolic and diastolic blood pressures and the insertion/deletion polymorphism of Angiotensin-converting enzyme, ACE and 2 protein polymorphisms of AGT (viz., M235T and T174M).

The genotypes of codon 174 were significantly associated with variation in systolic blood pressure in men and accounted for 3.1% of the total variation. Homozygotes for the AGT174M had the highest mean BP, followed by heterozygotes and homozygotes for AGT174T had the lowest mean BP (Hegele et al. 1996).

SNPs as Genetic Markers:

Most SNPs are not responsible for a disease state. Instead, they may serve as biological markers for tracing a disease gene(s) on the human genome map. Since SNPs occur frequently throughout the genome and is relatively stable, they serve as excellent biological markers. Biological markers are DNA segments with a pre-identified physical location in the chromosome, which can be easily tracked and used for constructing a chromosomal map of position of known genes relative to each other.

These maps allow the study identification of traits resulting from the interaction of more than one gene. Hence this strategy plays a major role in cases of complex gene

disorders. SNP markers, although biallelic, are preferred over the microsatellite markers as recurrent mutations are generally very rare in case of SNPs.

The National Centre for Biotechnology Information (NCBI) plays an important role in facilitating the identification and cataloguing of SNPs through the creation and maintenance of the public SNP database (dbSNP). This may be accessed by the biomedical community worldwide and is intended to facilitate many areas of biological research.

SNP in Linkage Disequilibrium Studies:

Particular alleles at neighbouring loci tend to be co-inherited. For tightly linked loci, this might lead to association between alleles in the population—a property known as linkage disequilibrium (LD) (Ardlie et al. 2002). The phenomenon of LD can be explained on the basis of co-segregation of two tightly linked alleles in a population, where one form of the haplotype is selected when the population experiences a bottleneck.

Later the selected haplotype becomes the founder haplotype as shown in (Fig. 20.4). Mutation and recombination have the most evident impact on LD. Additional factors contributing to the extent and distribution of LD are genetic drift, population growth, population admixture, migration, natural selection, variable recombination and mutation rates and gene conversion.

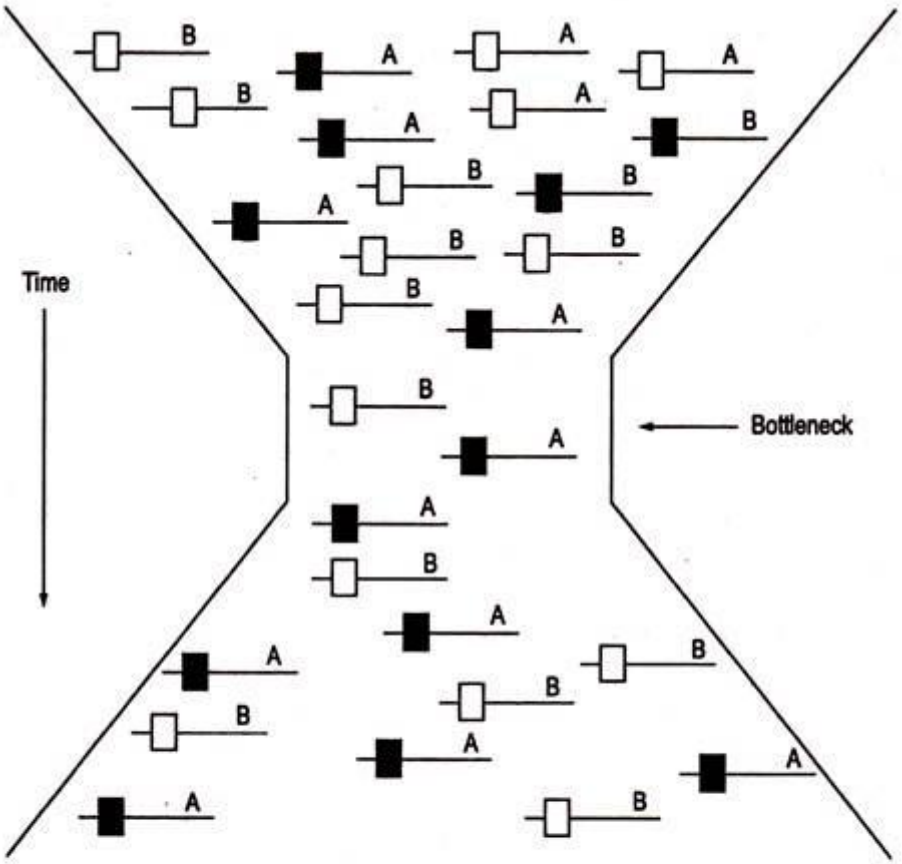


Fig. 20-4 : *A population bottleneck establishes linkage disequilibrium between a SNP and a disease allele. The illustration shows a representation of a population in a bottleneck. In the original population neither a disease allele (black-square) nor a normal allele (open-square) shows any particular association with a nearby SNP (i.e. A-allele or B-allele). However, the reduction in diversity in the bottleneck could result in loss or drastic reduction of haplotypes, as a result A-allele would be associated only with the disease allele (black-square) allowing the disease allele to be mapped (Sudbery, 2002)*

Measures of LD:

The Linkage Disequilibrium between Two Points A and B can be Calculated by the Expression:

$D = P_{AB} - P_A \infty P_B$

Where P_{AB} is the frequency of the haplotype that consists of allele A and B

P_A and P_B are the frequencies of the alleles A and B at loci A and B, respectively.

LD erosion occurs over time and distance. Hence the factor 'time' and 'distance' should be taken into consideration for calculation of LD.

If D_0 is the extent of disequilibrium at a starting point between two alleles, r distant apart, the disequilibrium t generations later (D_t):

$$D_t = (1- r)_t D_0$$

Complex Diseases and SNP:

Most of the genes responsible for major monogenic disorders have been mapped by positional cloning. These disorders follow the Mendelian pattern of inheritance. Diseases such as diabetes, cancer, asthma, rheumatoid arthritis do not show any clear pattern of such inheritance. Such disorders are referred to as complex or multifactorial diseases.

It is hypothesized that single nucleotide polymorphisms can be used for tracking genes responsible for complex disorders. For that purpose, SNPs which show co-segregation with a certain disease can be used as markers to identify (map) the loci responsible for the disease.

The identified SNPs in candidate genes for a complex disease could be used to determine susceptibility of an individual towards the disease, and when affected his SNP profile for genes related to drug target and drug metabolism could be used to determine the efficacy of the available drugs for therapeutic purposes.

The International HapMap Project: Understanding the Common Human Genetic Variations:

As described earlier, complex interaction of multiple genes, environmental factors and lifestyle result in common diseases, such as diabetes, cancer, stroke, cardiac diseases, psychiatric disorders, asthma etc. Although any two unrelated individuals are same at about 99.9% of their DNA sequences, the remaining 0.1% is important because it contains the genetic variants that provide their unique identity and also influence variability in their risk of disease and response to drugs. Discovering the DNA sequence variants that contribute to common disease risk offers one of the best opportunities for understanding the complex causes of disease in humans.

A centralized as well as multinational effort was required to map, discover and validate the common variations in the human genome. The first step towards grasping this knowledge was realized in 2001, with the completion of the human genome sequence. Consequently the International HapMap Project was initiated

The goal of the Project is to develop a haplotype map of the human genome, the HapMap, which will describe the common patterns of human DNA sequence variation. The Project is a powerful tool and a database, intended to facilitate the discovery of genetic contributions to common diseases, and can be used in studies that compare the patterns of genetic variation (haplotypes) in people with a specific disease to patterns in people without the disease.

By identifying regions of the human genome that shows differences in the haplotype patterns, particular genetic variants that contribute to the disease can be easily identified. To produce the HapMap, researchers analyzed blood samples from a total of 269 people from four large populations.These populations are: Yoruba in Ibadan, Nigeria, Japanese in Tokyo; Han Chinese in Beijing; and Utah residents with ancestry

from northen and western Europe. These four populations were selected to include people with ancestry from widely separate geographic regions—Caucasian (European and North American), Yoruban (Negroid) and Chinese and Japanese (Mongoloid). Interestingly, the Indian or the South Asian population is not represented in the HapMap.

In India several initiatives have been taken to study the genetic variations in the Indian population, the major one being 'The Indian Genome Variation database (IGVdb)'. The HapMap and the IGVdb projects have the potential to unravel the genetic basis of complex diseases leading to discoveries for prevention and treatment of such diseases.

An Indian Initiative:

The Indian subcontinent being a melting pot of different population and culture since the dawn of civilizations, the population can serve as an ideal model to study Single Nucleotide profile and its variety due to its diversity and ancient lineage, The Indian population comprises of more than a billion people, consisting of 4693 communities with thousands of endogamous groups, 325 functioning languages and 25 scripts.

To understand the origin, evolution, diversity and the migration patterns of the population, SNP serve as an ideal genetic tool. Furthermore, predisposition to complex disorders, variable sensitivity and reaction to different drugs is of prime importance. For this purpose, six constituent laboratories of the Council of Scientific and Industrial Research (CSIR) in collaboration with other premier research institutes of India initiated a network program.

The Indian Genome Variation (IGV) consortium to identify and validate SNPs and polymorphic repeat sequences in thousands of genes of the human genome. These genes have been selected on the basis of their relevance as functional and positional candidates in many common diseases including genes relevant to pharmacogenomics.

A review on the planned study is published in Human Genetics (The Indian Genome Variation database IGVdb, 2005). This is the first large-scale comprehensive effort from India to understand and utilize the already present genome variations for their deployment in the drug industry.

The blood samples on which the DNA analysis is to be done are being collected from multiple indigenous tribes and populations of India. The data is expected to give an insight to the Indian population structure, its evolution with a far reaching implication in the study of common complex diseases and pharmacogenomics.

Molecular phylogenetics:

Molecular phylogenetics is the branch of phylogeny that analyzes genetic, hereditary molecular differences, predominantly in DNA sequences, to gain information on an organism's evolutionary relationships. From these analyses, it is possible to determine the processes by which diversity among species has been achieved. The result of a molecular phylogenetic analysis is expressed in a phylogenetic tree. Molecular phylogenetics is one aspect of molecular systematics, a broader term that also includes the use of molecular data in taxonomy and biogeography.

Molecular phylogenetics and molecular evolution correlate. Molecular evolution is the process of selective changes (mutations) at a molecular level (genes, proteins, etc.) throughout various branches in the tree of life (evolution). Molecular phylogenetics makes inferences of the evolutionary relationships that arise due to molecular evolution and results in the construction of a phylogenetic tree.

History:

The theoretical frameworks for molecular systematics were laid in the 1960s in the works of Emile Zuckerkandl, Emanuel Margoliash, Linus Pauling, and Walter M. Fitch. Applications of molecular systematics were pioneered by Charles G. Sibley (birds), Herbert C. Dessauer (herpetology), and Morris Goodman (primates), followed by Allan C. Wilson, Robert K. Selander, and John C. Avise (who studied various groups). Work with protein electrophoresis began around 1956. Although the results were not quantitative and did not initially improve on morphological classification, they provided tantalizing hints that long-held notions of the classifications of birds, for example, needed substantial revision. In the period of 1974–1986, DNA-DNA hybridization was the dominant technique used to measure genetic difference.

Theoretical background :

Early attempts at molecular systematics were also termed as chemotaxonomy and made use of proteins, enzymes, carbohydrates, and other molecules that were separated and characterized using techniques such as chromatography. These have been replaced in recent times largely by DNA sequencing, which produces the exact sequences of nucleotides or bases in either DNA or RNA segments extracted using different techniques. In general, these are considered superior for evolutionary studies, since the actions of evolution are ultimately reflected in the genetic sequences. At present, it is still a long and expensive process to sequence the entire DNA of an organism (its genome). However, it is quite feasible to determine the sequence of a defined area of a particular chromosome. Typical molecular systematic analyses require the sequencing of around 1000 base pairs. At any location within such a sequence, the bases found in a given position may vary between organisms. The particular sequence found in a given organism is referred to as its haplotype. In principle, since there are four base types, with 1000 base pairs, we could have 41000 distinct haplotypes. However, for organisms

within a particular species or in a group of related species, it has been found empirically that only a minority of sites show any variation at all, and most of the variations that are found are correlated, so that the number of distinct haplotypes that are found is relatively small.

In a molecular systematic analysis, the haplotypes are determined for a defined area of genetic material; a substantial sample of individuals of the target species or other taxon is used; however, many current studies are based on single individuals. Haplotypes of individuals of closely related, yet different, taxa are also determined. Finally, haplotypes from a smaller number of individuals from a definitely different taxon are determined: these are referred to as an outgroup. The base sequences for the haplotypes are then compared. In the simplest case, the difference between two haplotypes is assessed by counting the number of locations where they have different bases: this is referred to as the number of substitutions (other kinds of differences between haplotypes can also occur, for example, the insertion of a section of nucleic acid in one haplotype that is not present in another). The difference between organisms is usually re-expressed as a percentage divergence, by dividing the number of substitutions by the number of base pairs analysed: the hope is that this measure will be independent of the location and length of the section of DNA that is sequenced.

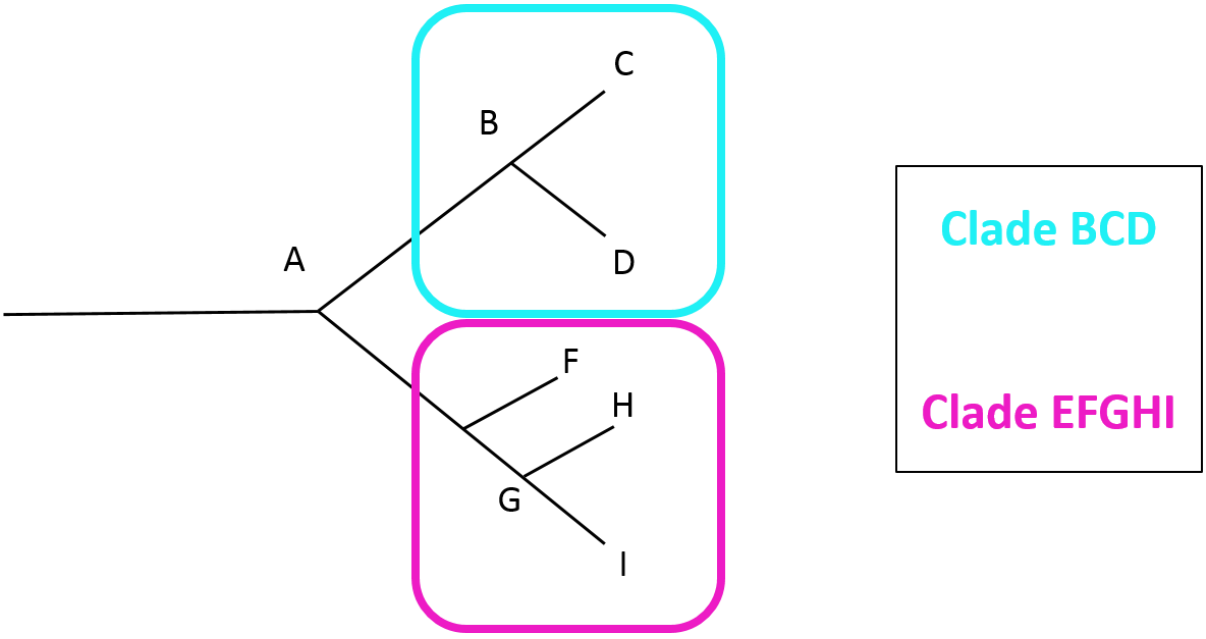


Figure: In a phylogenetic tree, numerous groupings (clades) exist. A clade may be defined as a group of organisms having a common ancestor throughout evolution. This figure illustrates how a clade in a phylogenetic tree may be expressed.

An older and superseded approach was to determine the divergences between the genotypes of individuals by DNA-DNA hybridization. The advantage claimed for using hybridization rather than gene sequencing was that it was based on the entire genotype, rather than on particular sections of DNA. Modern sequence comparison techniques overcome this objection by the use of multiple sequences.

Once the divergences between all pairs of samples have been determined, the resulting triangular matrix of differences is submitted to some form of statistical cluster analysis, and the resulting dendrogram is examined in order to see whether the samples cluster in the way that would be expected from current ideas about the taxonomy of the group. Any group of haplotypes that are all more similar to one another than any of them is to any other haplotype may be said to constitute a clade, which may be visually represented as the figure displayed on the right demonstrates. Statistical techniques such as bootstrapping and jackknifing help in providing reliability estimates for the positions of haplotypes within the evolutionary trees.'

Techniques and applications:

Every living organism contains deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins. In general, closely related organisms have a high degree of similarity in the molecular structure of these substances, while the molecules of organisms distantly related often show a pattern of dissimilarity. Conserved sequences, such as mitochondrial DNA, are expected to accumulate mutations over time, and assuming a constant rate of mutation, provide a molecular clock for dating divergence. Molecular phylogeny uses such data to build a "relationship tree" that shows the probable evolution of various organisms. With the invention of Sanger sequencing in 1977, it became possible to isolate and identify these molecular structures. High-throughput sequencing may also be used to obtain the transcriptome of an organism, allowing inference of phylogenetic relationships using transcriptomic data.

The most common approach is the comparison of homologous sequences for genes using sequence alignment techniques to identify similarity. Another application of molecular phylogeny is in DNA barcoding, wherein the species of an individual organism is identified using small sections of mitochondrial DNA or chloroplast DNA. Another application of the techniques that make this possible can be seen in the very limited field of human genetics, such as the ever-more-popular use of genetic testing to determine a child's paternity, as well as the emergence of a new branch of criminal forensics focused on evidence known as genetic fingerprinting.

Molecular phylogenetic analysis:

There are several methods available for performing a molecular phylogenetic analysis. One method, including a comprehensive step-by-step protocol on constructing a phylogenetic tree, including DNA/Amino Acid contiguous sequence assembly, multiple sequence alignment, model-test (testing best-fitting substitution models), and phylogeny reconstruction using Maximum Likelihood and Bayesian Inference, is available at Nature Protocol.

Another molecular phylogenetic analysis technique has been described by Pevsner and shall be summarized in the sentences to follow (Pevsner, 2015). A phylogenetic analysis

typically consists of five major steps. The first stage comprises sequence acquisition. The following step consists of performing a multiple sequence alignment, which is the fundamental basis of constructing a phylogenetic tree. The third stage includes different models of DNA and amino acid substitution. Several models of substitution exist. A few examples include Hamming distance, the Jukes and Cantor one-parameter model, and the Kimura two-parameter model. The fourth stage consists of various methods of tree building, including distance-based and character-based methods. The normalized Hamming distance and the Jukes-Cantor correction formulas provide the degree of divergence and the probability that a nucleotide changes to another, respectively. Common tree-building methods include unweighted pair group method using arithmetic mean (UPGMA) and Neighbor joining, which are distance-based methods, Maximum parsimony, which is a character-based method, and Maximum likelihood estimation and Bayesian inference, which are character-based/model-based methods. UPGMA is a simple method; however, it is less accurate than the neighbor-joining approach. Finally, the last step comprises evaluating the trees. This assessment of accuracy is composed of consistency, efficiency, and robustness.

MEGA (molecular evolutionary genetics analysis) is an analysis software that is user-friendly and free to download and use. This software is capable of analyzing both distance-based and character-based tree methodologies. MEGA also contains several options one may choose to utilize, such as heuristic approaches and bootstrapping. Bootstrapping is an approach that is commonly used to measure the robustness of topology in a phylogenetic tree, which demonstrates the percentage each clade is supported after numerous replicates. In general, a value greater than 70% is considered significant. The flow chart displayed on the right visually demonstrates the order of the five stages of Pevsner's molecular phylogenetic analysis technique that have been described.[13]

Fig: Five Stages of Molecular Phylogenetic Analysis

Limitations:

Molecular systematics is an essentially cladistic approach: it assumes that classification must correspond to phylogenetic descent, and that all valid taxa must be monophyletic. This is a limitation when attempting to determine the optimal tree(s), which often involves bisecting and reconnecting portions of the phylogenetic tree(s).

The recent discovery of extensive horizontal gene transfer among organisms provides a significant complication to molecular systematics, indicating that different genes within the same organism can have different phylogenies.In addition, molecular phylogenies are sensitive to the assumptions and models that go into making them. Firstly, sequences must be aligned; then, issues such as long-branch attraction, saturation, and

taxon sampling problems must be addressed. This means that strikingly different results can be obtained by applying different models to the same dataset.

Moreover, as previously mentioned, UPGMA is a simple approach in which the tree is always rooted. The algorithm assumes a constant molecular clock for sequences in the tree. This is associated with being a limitation in that if unequal substitution rates exist, the result may be an incorrect tree.

Probable Questions:

1. Define single nucleotide polymorphism.
2. Differentiate transition and transversion.
3. Discuss genetic predisposition.
4. State applications of SNP in pharmacogenomics.
5. How SNPs in promoter affects gene expression.
6. How SNPs are used as genetic markers?
7. Sate limitations of molecular systematic.
8. What is phylogenetic trees? How it is constructed?
9. Define molecular taxonomy. How it is prepared?

Suggested Readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics . Verma and Agarwal

UNIT-III

Genetics in forensic science: DNA comparisons, Genetic finger-printing, RFLPs, AFLPS, VNTRs and Proteomics.

Objective: In this unit we will discuss about different methods of DNA Fingerprinting, RFLPs, AFLPs, VNTRs, and alsosome basic ideas about proteomics.

Meaning of DNA Polymorphisms:

Different alleles of a gene produce different phenotypes which can be detected by making crosses between parents with different alleles of two or more genes. Then by determining recombinants in the progeny, a genetic map can be deduced.These are low resolution genetic maps that contain genes with observable phenotypic effects, all mapped to their respective loci. The position of a specific gene, or locus can be found from the map. However, measurements showed that the chromosomal intervals between the mapped genes would contain vast amounts of DNA.

These intervals could not be mapped by the recombinant progeny method because there were no markers in those intervening regions. It became necessary to find additional differential markers or genetic differences that fall in the gaps. This need was met by exploitation of various polymorphic DNA markers.A DNA polymorphism is a DNA sequence variation that is not associated with any observable phenotypic variation, and can exist anywhere in the genome, not necessarily in a gene. Polymorphism means one of two or more alternative forms (alleles) of a chromosomal region that either has a different nucleotide sequence, or it has variable numbers of tandemly repeated nucleotides.

Thus, it is a site of heterozygosity for any sequence variation. Many DNA polymorphisms are useful for genetic mapping studies, hence they are referred to as DNA markers. DNA markers can be detected on Southern blot hybridisation or by PCR. The alleles of DNA markers are co-dominant, that is they are neither dominant nor recessive as observed in alleles of most genes. DNA polymorphisms constitute molecularly defined differences between individual human beings.

DNA Fingerprinting:

DNA fingerprinting is the present day genetic detective in the practice of modern medical forensics. The underlying principles of DNA fingerprinting are briefly described. The structure of each person's genome is unique. The only exception being monozygotic

identical twins (twins developed from a single fertilized ovum).The unique nature of genome structure provides a good opportunity for the specific identification of an individual. It may be remembered here that in the traditional fingerprint technique, the individual is identified by preparing an ink impression of the skin folds at the tip of the person's finger. This is based on the fact that the nature of these skin folds is genetically determined, and thus the fingerprint is unique for an individual. In contrast, the DNA fingerprint is an analysis of the nitrogenous base sequence in the DNA of an individual.

History and Terminology:

The original DNA fingerprinting technique was developed by Alec Jaffrey's in 1985. Although the DNA fingerprinting is commonly used, a more general term DNA profiling is preferred. This is due to the fact that a wide range of tests can be carried out by DNA sequencing with improved technology.

Applications of DNA Fingerprinting:

The amount of DNA required for DNA fingerprint is remarkably small. The minute quantities of DNA from blood strains, body fluids, and hair fiber or skin fragments are enough. Polymerase chain reaction is used to amplify this DNA for use in fingerprinting. DNA profiling has wide range of applications—most of them related to medical forensics.

Some important ones are listed below:

i. Identification of criminals, rapists, thieves etc.

ii. Settlement of paternity disputes.

iii. Use in immigration test cases and disputes.

In general, the fingerprinting technique is carried out by collecting the DNA from a suspect (or a person in a paternity or immigration dispute) and matching it with that of a reference sample (from the victim of a crime, or a close relative in a civil case).

DNA Markers in Disease Diagnosis and Fingerprinting:

The DNA markers are highly useful for genetic mapping of genomes. There are four types of DNA sequences which can be used as markers.

1. Restriction fragment length polymorphisms (RFLF).

2. Minisatellites or variable number tandem repeats (VNTR).

3. Microsatellites or simple tandem repeats (STRs).

4. Single nucleotide polymorphisms (SNPs, pronounced as snips).

The general aspects of the above DNA markers are described along with their utility in disease diagnosis and DNA fingerprinting.

Restriction Fragment Length Polymorphisms (RFLPs):

A RFLP represents a stretch of DNA that serves as a marker for mapping a specified gene. RFLPs are located randomly throughout a person's chromosomes and have no apparent function. A DNA molecule can be cut into different fragments by a group of enzymes called restriction endonucleases .These fragments are called polymorphisms (literally means many forms).

An outline of RFLP is depicted in Fig. 14.5. The DNA molecule 1 has three restriction sites (R_1, R_2, R_3), and when cleaved by restriction endonucleases forms 4 fragments. Let us now consider DNA 2 with an inherited mutation (or a genetic change) that has altered some base pairs. As a result, the site (R_2) for the recognition by restriction endonuclease is lost. This DNA molecule 2 when cut by restriction endonuclease forms only 3 fragments (instead of 4 in DNA 1).

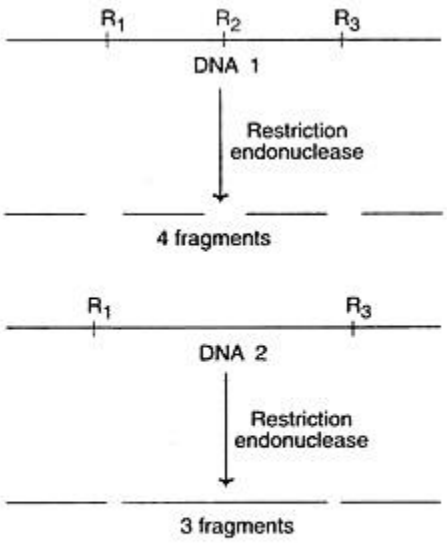


***Fig. 14.5 :** An outline of the restriction fragment length polymorphism (RFLP) (R_1, R_2, R_3 represent the sites for the action of restriction endonucleases).*

As is evident from the above description, a stretch of DNA exists in fragments of various lengths (polymorphisms), derived by the action of restriction enzymes, hence the name restriction fragment length polymorphisms.

RFLPs in the Diagnosis of Diseases:

If the RFLP lies within or even close to the locus of a gene that causes a particular disease, it is possible to trace the defective gene by the analysis of RFLP in DNA. The

person's cellular DNA is isolated and treated with restriction enzymes. The DNA fragments so obtained are separated by electrophoresis.

The RFLP patterns of the disease suspected individuals can be compared with that of normal people (preferably with the relatives in the same family). By this approach, it is possible to determine whether the individual has the marker RFLP and the disease gene. With 95% certainty, RFLPs can detect single gene-based diseases.

Methods of RFLP scoring:

Two methods are in common use for the detection of RFLPs (Fig. 14.5).

1. Southern hybridization:

The DNA is digested with appropriate restriction enzyme, and separated by agarose gel electrophoresis. The so obtained DNA fragments are transferred to a nylon membrane. A DNA probe that spans the suspected restriction site is now added, and the hybridized bands are detected by autoradiograph. If the restriction site is absent, then only a single restriction fragment is detected. If the site is present, then two fragments are detected (Fig. 14.6A).

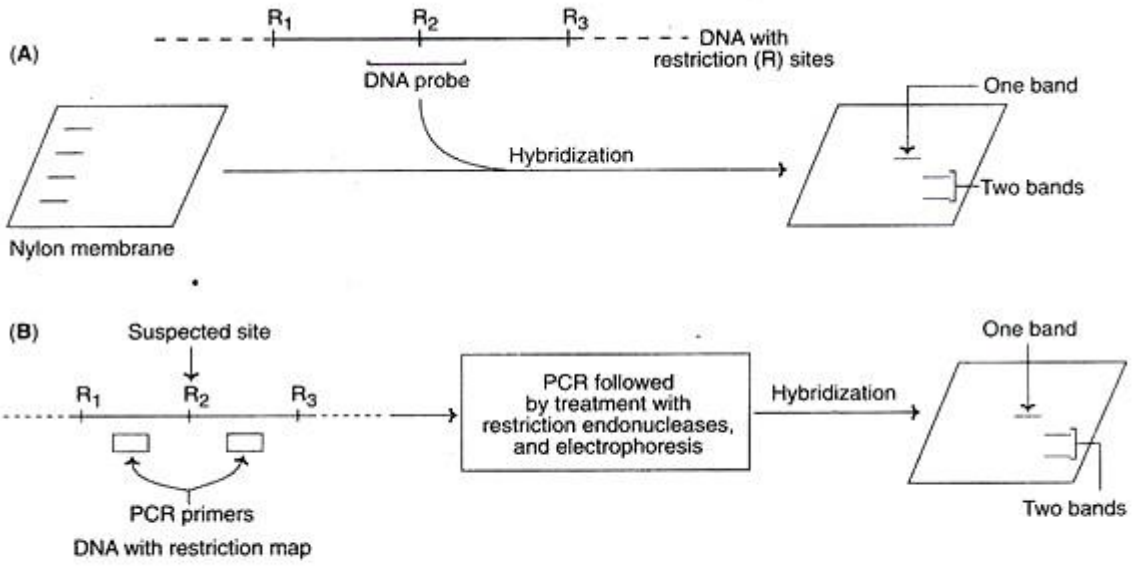


Fig. 14.6 : Two common methods used for scoring restriction fragment length polymorphism (RFLP) (A) RFLP by Southern hybridization (B) RFLP by polymerase chain reaction (PCR).

2. Polymerase chain reaction:

RFLPs can also be scored by PCR. For this purpose, PCR primers that can anneal on either side of the suspected restriction site are used. After amplification by PCR, the DNA molecules are treated with restriction enzyme and then analysed by agarose gel

electrophoresis. If the restriction site is absent only one band is seen while two bands are found if the site is found (Fig. 14.6B).

Applications of RFLPs:

The approach by RFLP is very powerful and has helped many genes to be mapped on the chromosomes, e.g. sickle- cell anaemia (chromosome 11), cystic fibrosis (chromosome 7), Huntington's disease (chromosome 4), retinoblastoma (chromosome 13), Alzheimer's disease (chromosome 21)

Variable Number Tandem Repeats (VNTRs):

VNTRs, also known as mini-satellites, like RFLPs, are DNA fragments of different length. The main difference is that RFLPs develop from random mutations at the site of restriction enzyme activity while VNTRs are formed due to different number of base sequences between two points of a DNA molecule. In general, VNTRs are made up of tandem repeats of short base sequences (10-100 base pairs). The number of elements in a given region may vary, hence they are known as variable number tandem repeats.

An individual's genome has many different VNTRs and RFLPs which are unique to the individual. The pattern of VNTRs and RFLPs forms the basis of DNA fingerprinting or DNA profiling. In the Fig. 14.7, two different DNA molecules with different number of copies (bands) of VNTRs are shown. When these molecules are subjected to restriction endonuclease action (at two sites R_1 and R_2), the VNTR sequences are released, and they can be detected due to variability in repeat sequence copies. These can be used in mapping of genomes, besides their utility in DNA fingerprinting.

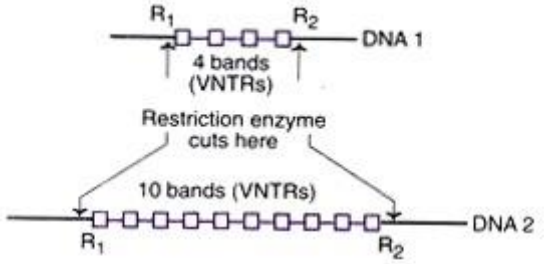


Fig. 14.7 : A diagrammatic representation of variable number tandem repeats (VNTRs). Each band (or copy) represents a repeating sequence in the DNA (e.g. 100 base pairs each). R_1 and R_2 indicate the sites cut by a restriction enzyme.

VNTRs are useful for the detection of certain genetic diseases associated with alterations in the degree of repetition of microsatellites e.g. Huntington's chorea is a disorder which is found when the VNTRs exceed 40 repeat units.

Limitations of VNTRs:

The major drawback of VNTRs is that they are not evenly distributed throughout the genome. VNTRs tend to be localized in the telomeric regions at the ends of the chromosomes.

Use of RFLPs and VNTRs in Genetic Fingerprinting:

RFLPs caused by variations in the number of VNTRs between two restriction sites can be detected (Fig. 14.8). The DNAs from three individuals with different VNTRs are cut by the specific restriction endonuclease. The DNA fragments are separated by electrophoresis, and identified after hybridization with a probe complementary to a specific sequence on the fragments.

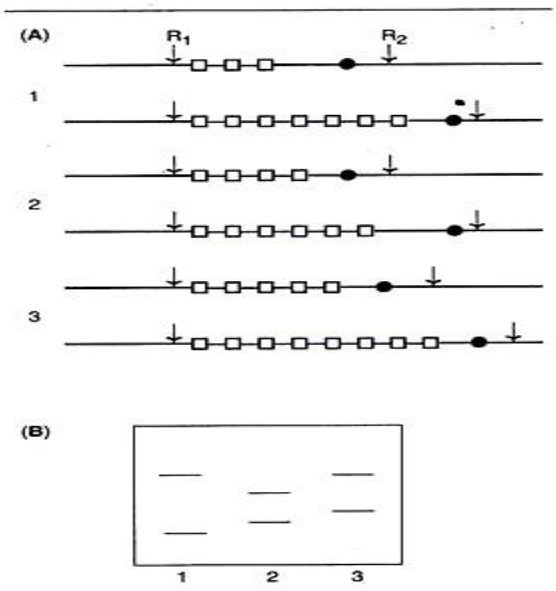


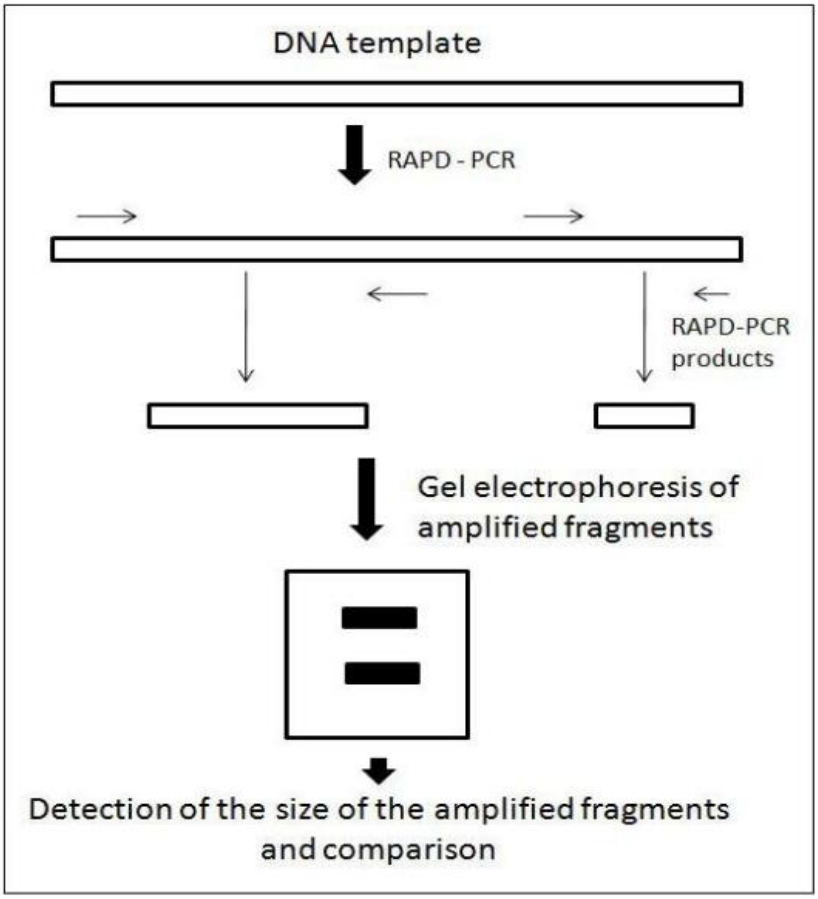
Fig. 14.8 : Use of restriction fragment length polymorphisms (RFLPs) caused by variable number tandem repeats (VNTRs) in genetic fingerprinting (A) An illustration of DNA structure from three individuals (B) Hybridized pattern of DNA fragment with a probe complementary to the sequence shown in black circles (1, 2 and 3 represent the individuals; R_1 and R_2 indicate restriction sites; coloured squares are the number of VNTRs)

Randomly Amplified Polymorphic DNA:

RAPDs are based on random PCR amplification. The procedure is carried out by randomly designing primers for PCR which will amplify several different regions of the genome by chance. Such a primer results in amplification of only those DNA regions that have near them, inverted copies of the primer's own sequence.

The PCR products consist of DNA bands representing different sizes of the amplified DNA. The set of amplified DNA fragments is called randomly amplified polymorphic

DNA (RAPD). Certain bands may be unique for an individual and can serve as DNA markers in mapping analysis.



Amplified fragment length polymorphism (AFLP):

AFLP is a novel technique involving a combination of RFLP and RAPD. AFLP is based on the principle of generation of DNA fragments using restriction enzymes and oligonucleotide adaptors (or linkers), and their amplification by PCR. Thus, this technique combines the usefulness of restriction digestion and PCR.

The DNA of the genome is extracted. It is subjected to restriction digestion by two enzymes (a rare cutter e.g. MseI; a frequent cutter e.g. EcoRI). The cut ends on both sides are then ligated to known sequences of oligonucleotides (Fig. 53.5).

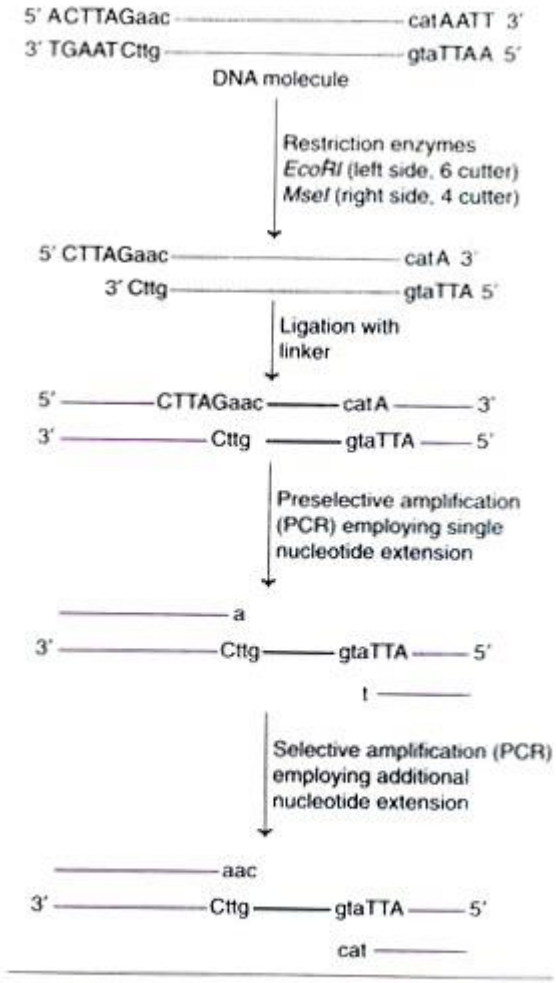


***Fig. 53.5** : A diagrammatic representation of the amplified fragment length polymorphism (AFLP) (**Note** : The lower case letters represent the sequences found within the amplified region; the coloured lines indicate linkers).*

PCR is now performed for the pre-selection of a fragment of DNA which has a single specific nucleotide. By this approach of pre-selective amplification, the pool of fragments can be reduced from the original mixture. In the second round of amplification by PCR, three nucleotide sequences are amplified.

This further reduces the pool of DNA fragments to a manageable level (< 100). Autoradiography can be performed for the detection of DNA fragments. Use of radiolabelled primers and fluorescently labelled fragments quickens AFLP.AFLP analysis is tedious and requires the involvement of skilled technical personnel. Hence some people are not in favour of this technique. In recent years, commercial kits are made available for AFLP analysis. AFLP is very sensitive and reproducible. It does not require prior knowledge of sequence information. By AFLP, a large number of polymorphic bands can be produced and detected.

Microsatellites (Simple Tandem Repeats):

Microsatellites are short repeat units (10-30 copies) usually composed of dinucleotide or tetra nucleotide units. These simple tandem repeats (STRs) are more popular than mini-satellites (VNTRs) as DNA markers for two reasons.

1. Microsatellites are throughout the genome.

2. PCR can be effectively and conveniently used to identify the length of polymorphism.

Two variants (alleles) of DNA molecules with 5 and 10 repeating units of a dimer nucleotides (GA) are depicted in Fig. 14.9.

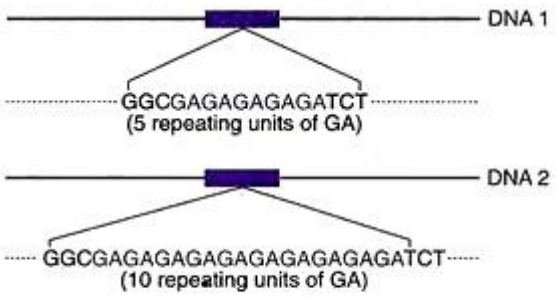


Fig. 14.9 : Two alleles of DNA molecules representing 5 and 10 dimer repeating units.

By use of PCR, the region surrounding the microsatellites is amplified, separated by agarose gel electrophoresis and identified.

Single Nucleotide Polymorphisms (SNPs):

SNPs represent the positions in the genome where some individuals have one nucleotide (e.g. G) while others have a different nucleotide (e.g. C). There are large numbers of SNPs in genomes. It is estimated that the human genome contains at least 3 million SNPs. Some of these SNPs may give rise to RFLPs. SNPs are highly useful as DNA markers since there is no need for gel electrophoresis and this saves a lot of time and labour. The detection of SNPs is based on the oligonucleotide hybridization analysis (Fig. 14.10).

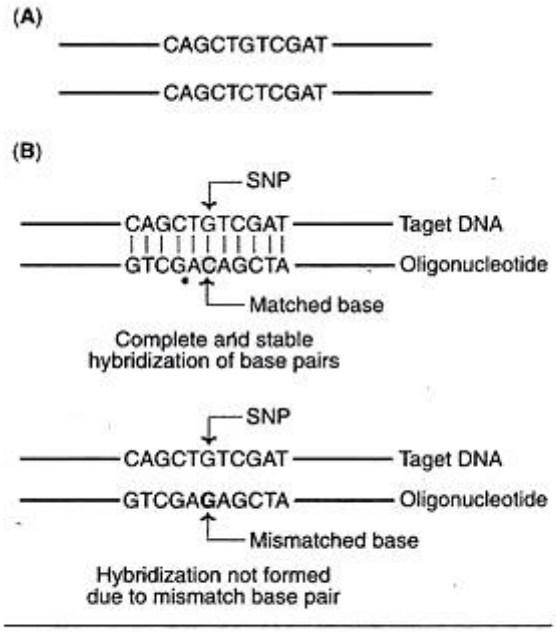


Fig. 14.10 : (A) An illustration of single nucleotide polymorphism (SNP) (B) Oligonucleotide hybridization to detect SNP.

An oligonucleotide is a short single-stranded DNA molecule synthesized in the laboratory with a length not usually exceeding 50 nucleotides. Under appropriate conditions, this nucleotide sequence will hybridize with a target DNA strand if both have completely base paired structure. Even a single mismatch in base pair will not allow the hybridization to occur. DNA chip technology is most commonly used to screen SNPs hybridization with oligonucleotide. About one-half of missense mutations that are SNPs are estimated to cause genetic disease in humans. A non-coding SNP can also affect gene function if it is located in the promoter region or in the gene regulatory region. A small number of SNPs can create a restriction site, or eliminate an already existing restriction site. SNP-induced alterations in restriction sites are detected by using the restriction enzyme followed by Southern blot analysis or PCR.

An individual SNP locus can be analysed by using the technique of allele-specific oligonucleotide (ASO) hybridisation. The search for one particular SNP locus in humans is a challenge, because this is one base pair that is polymorphic out of the three billion base pairs in the human genome.In the ASO technique, a short oligonucleotide that is complementary to one SNP allele is synthesised and mixed with the target DNA. Hybridisation is performed under high stringency conditions that would allow only a perfect match between probe and the target DNA. That means, the oligonucleotide will not hybridize with target DNA that has any other SNP allele at that locus. Positive result of hybridisation indicates the SNP locus precisely. A more recent technique of DNA Microarrays can be used for simultaneous typing of hundreds or thousands of SNPs. Details of this technique used for SNPs and genome wide gene expression are described later in this section.

Current Technology of DNA Fingerprinting:

In the forensic analysis of DNA, the original techniques based on RFLPs and VNTRs are now largely replaced by microsatellites (short tandem repeats). The basic principle involves the amplification of microsatellites by polymerase chain reaction followed by their detection. It is now possible to generate a DNA profile by automated DNA detection system (comparable to the DNA sequencing

Genetic Profiling: DNA profiling (also called DNA fingerprinting) is the process of determining an individual's DNA characteristics, which are as unique as fingerprints. DNA analysis intended to identify a species, rather than an individual, is called DNA barcoding.

DNA profiling is a forensic technique in criminal investigations, comparing criminal suspects' profiles to DNA evidence so as to assess the likelihood of their involvement in the crime. It is also used in parentage testing, to establish immigration eligibility, and in genealogical and medical research. DNA profiling has also been used in the study of animal and plant populations in the fields of zoology, botany, and agriculture. Starting in the 1980s scientific advances allowed for the use of DNA as a mechanism for the identification of an individual. The first patent covering the modern process of DNA profiling was filed by Dr. Jeffrey Glassberg in 1983, based upon work he had done while at Rockefeller University in 1981. Glassberg, along with two medical doctors, founded Lifecodes Corporation to bring this invention to market. The Glassberg patent was issued in Belgium BE899027A1, Canada FR2541774A1, Germany DE3407196 A1, Great Britain GB8405107D0, Japan JPS59199000A, United States as US5593832A. In the United Kingdom, Geneticist Sir Alec Jeffreys independently developed a DNA profiling process in beginning in late 1984 while working in the Department of Genetics at the University of Leicester.

The process, developed by Jeffreys in conjunction with Peter Gill and Dave Werrett of the Forensic Science Service (FSS), was first used forensically in the solving of the murder of two teenagers who had been raped and murdered in Narborough, Leicestershire in 1983 and 1986. In the murder inquiry, led by Detective David Baker, the DNA contained within blood samples obtained voluntarily from around 5,000 local men who willingly assisted Leicestershire Constabulary with the investigation, resulted in the exoneration of Richard Buckland, an initial suspect who had confessed to one of the crimes, and the subsequent conviction of Colin Pitchfork on January 2, 1988. Pitchfork, a local bakery employee, had coerced his co-worker Ian Kelly to stand in for him when providing a blood sample—Kelly then used a forged passport to impersonate Pitchfork. Another co-worker reported the deception to the police. Pitchfork was arrested, and his blood was sent to Jeffrey's lab for processing and profile development. Pitchfork's profile matched that of DNA left by the murderer which confirmed Pitchfork's presence at both crime scenes; he pleaded guilty to both murders.

Although 99.9% of human DNA sequences are the same in every person, enough of the DNA is different that it is possible to distinguish one individual from another, unless they are monozygotic (identical) twins. DNA profiling uses repetitive sequences that are highly variable, called variable number tandem repeats (VNTRs), in particular short tandem repeats (STRs), also known as microsatellites, and minisatellites. VNTR loci are similar between closely related individuals, but are so variable that unrelated individuals are unlikely to have the same VNTRs.

In India DNA fingerprinting was started by Dr. VK Kashyap and Dr. Lalji Singh. Singh was an Indian scientist who worked in the field of DNA fingerprinting technology in India, where he was popularly known as the "Father of Indian DNA fingerprinting". In 2004, he received the Padma Shri in recognition of his contribution to Indian science and technology.

Proteomics:

The term proteomics was coined in mid 1990s at the back drop of successful genomics. In bioinformatics point of view proteomics is the databases of protein sequence, databases of predicted protein structures and more recently, databases of protein expression analysis. As more protein structures are identified, the relationship between structure and functions became easier to predict.

In addition, databases of protein structure and corporating tools facilitating the identification of common protein structure and their predicted functions. In this technique individually purified ligands such as proteins, peptides, antibodies, antigens, and carbohydrates are spotted on to a derivatized surface and are generally used for examining protein expression levels for protein profiling. A major challenge facing plant biotechnology and other bioinformatics research community is the translation of complete genome DNA sequence data into protein structure and predicted functions. Such a steps will provide the key link between the genotypes of an organism and its expressed phenotype.

The growth of proteomics is a direct result of advances made in large scale nucleotide sequencing of expressed sequence tags (EST). Although mass spectrometry or more popularly MS technology has been considered as versatile tool for examining simultaneous expression of more than 1000 proteins and identification, mapping of post-translational modifications (Table 25.5).These methods performed in a latest array of technology resulted in large-scale characterization of protein location, protein-protein interaction and protein functions.

Table 25.5 Proteomics tools

Method	Description	Applications
1. Mass spectrophotometer	Digest protein and fragment peptide to identification proteins,	Protein identification, sequence post translational modification
2. Chip	Synthesise proteins, peptides, antigens, antibody into a avery format and spot onto slides	Protein interaction with protein, lipid and small molecules, drug discover, post translational modifications.
3. Bioinformatics	Insilico proteomics	Mining database predicting protein interaction.

In-silico methodologies are being developed to identify protein interaction from genome sequence. For example, 6809 putative protein-protein interaction has been identified in Escherichia coli and more than 45,000 have been identified in yeast and large number of these interactions is functionally related.

Types of Proteomics:

i. Structural Proteomics:

One of the main targets of proteomics investigation is to map the structure of protein complexes or the proteins present in a specific cellular organelle known as cell map or structural proteins. Structural proteomics attempt to identify all the proteins within a protein complex and characterization all protein-protein interactions. Isolation of specific protein complex by purification can simplify the proteomic analysis.

ii. Functional Proteomics:

It mainly includes isolation of protein complexes or the use of protein ligands to isolate specific types of proteins. It allows selected groups of proteins to be studied its characteristics which can provide important information about protein signalling and disease mechanism etc.

Significance of Proteomics:

i. Protein profiling:

Bioinformatics has been widely employed in protein-profiling, where question of protein structural information for the purpose of protein identification, characterization and database is carried out. The spectrum of protein expressed in a cell type provides the cell with its unique identity. It explores how the protein complement changes in a cell type during development in response to environmental stress.

ii. Protein arrays:

Protein microarrays facilitate the detection of protein protein interaction and protein expression profiling. Several protein microarray examples indicate that protein arrays hold great promise for the global analysis of protein-protein and protein-ligand interaction.

iii. Proteomics to a phosphorylation:

In post-translational modification of protein, mass spectrometer (MS) can be used to identify novel phosphorylation. Measure changes in phosphorylation state of protein takes place in response to an effective and determining phosphorylation sites in proteins.

Identification of phosphorylation sites can provide information about the mechanism of enzyme regulation and protein kinase and phosphotases involved. A proteomics approach for this process has an advantage that one can study all the phosphorylating proteins in a cell at the same time.

iv. Proteome mining:

Proteome mining is a functional proteomic approach used to extract information from the analysis of specific sub-proteomics. In principle, it is based on the assumption. In principle, it is based oil the assumption that all drug like molecule selectively compete with a natural cellular ligand for a binding site on a protein target.

Basic Concepts of Proteomics:

The gene transcripts that an individual can make in a lifetime—termed as transcriptome (by analogy with the term genome)—refers to the haploid set of chromosomes carrying all the functional genes. Similarly, all the proteins made by an organism are now grouped under the shade of proteomics. Proteomics involves the systematic study of proteins in order to provide a comprehensive view of the structure, function and role in the regulation of a biological system.

These include protein-protein interaction, protein modification, protein function and its localization studies. The aim of proteomics is not only to identify all the proteins in a cell but also to create a complete three-dimensional map of the cell indicating where proteins are located. Coupled with advances in bioinformatics, this approach to comprehensively describing biological systems will undoubtedly have a major impact on our understanding of the phenotype of both normal and diseased cells. The proteome (term coined by Mark Wilkins in 1995) of a given cell is the total number of proteins at any given instant and it is highly dynamic in response to internal and

external cues. Proteins can be modified by post-translational modifications, undergo translocations within the cell or be synthesized or degraded.

Therefore, the examination of proteins of a cell at a particular time reflects the immediate protein environment in which it is studied. A cellular proteome is the collection of proteins found in a particular cell type under the influence of a particular set of environmental conditions like exposure to hormone stimulation. A complete set of proteins from all of the various cellular proteomes will form an organism's complete proteome. An interesting finding of the Human Genome Project is that there are far more proteins in the human proteome (~ 400,000 proteins) than there are protein-coding genes in the human genome (~ 22,000 genes). The large increase in protein diversity is thought to be due to alternative splicing and post-translational modification of proteins. This indicates that protein diversity cannot be fully characterized by gene expression analysis alone. Proteomics, thus is a useful tool for characterizing cells and tissues of interest.

The first protein studies that can be called proteomics began with the introduction of two dimensional gel electrophoresis of *E. coli* proteins (O'Ferrall, 1975) followed by mouse and guinea pig protein studies (Ksole, 1975). Although 2-dimensional electrophoresis (2-DE) was a major step forward and many proteins could be separated and visualized by this technique but it was not enough for the protein identification through any sensitive protein sequencing technology. After certain efforts the first major technology for the identification of protein was protein sequencing by Edman degradation (Edman, 1949). This technology was used for the identification of proteins from 2-D gels to create first 2D database (Celis et al. 1987). Another most important development in protein identification was Mass Spectrometry (MS) technology (Andersen et al. 2000). Protein sequencing by MS technology has been increased due to its sensitivity of analysis, tolerate protein complexes and amenable to high throughput operations.

Although several advancements have been made in protein identification (by MS or Edman sequencing) without having the database of large scale DNA sequencing of expressed sequences and genomic DNA, proteins could not be characterized because different protein isoforms can be generated from a single gene through several modifications (Fig. 18.1). And the majority of DNA and protein sequences have been accumulated within a short period of time.

In 1995, the sequencing of the genome of an organism was done for the first time in Haemophilus influenzae (Fleischmann et al. 1995). Till date, sequencing of several other eukaryotic genomes have been completed viz. Arabidopsis thaliana (Tabata, 2000), *Sachcharomyces cerevisiae* (Goffeau, 1996), Caenorhabditis elegans (Abbott, 1998), Oryza (Matsumoto, 2001) and human (Venter, 2001).

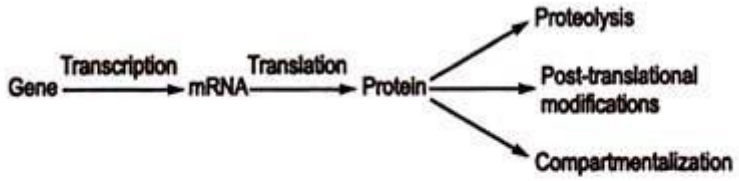


Fig. 18·1 : *Diagrammatic representation of a gene expression showing formation of many protein isoforms from a single gene. After transcription of the gene, mRNA is alternatively spliced or edited to form a mature mRNA that is translated to the protein. Proteins can be regulated by additional mechanism of proteolysis, compartmentalization and certain other modifications*

For protein expression profiling, a common procedure is the analysis of mRNA by different methods including serial analysis of gene expression (SAGE) (Velculescu et al. 1995) and DNA microarray technology (Shalon, 1996). However, the level of transcription of a gene gives only a rough idea of the real level of expression of that gene.

An mRNA may be produced in abundance, but at the same time degraded rapidly, or translated inefficiently keeping the amount of protein minimum. Proteins having been formed are subjected to post-translational modifications also. Different post-translational modifications or proteolysis and compartmentalization regulate the protein functions in the cell (Fig. 18.1).The average number of proteins formed per gene was predicted to be one or two in bacterium, three in yeast and three or more in humans (Wilkins et al. 1996). In response to extra-cellular responses, a number of proteins undergo post-translational modifications. Protein phosphorylation is an important signalling mechanism and dis-regulation of protein kinase and phosphatase can result oncogenesis (Hunter, 1995).

Through proteome analysis, changes in the modifications of many proteins expressed by a cell can be analysed after translation. Another important feature of a protein is its localization in the cell. The mis-localization of proteins is known to have an adverse effect on cellular function (cystic fibrosis) (Drumm and Collins, 1993). The cell growth, programmed cell death and the decision to proceed through the cell cycle are all regulated by signal transduction through protein complexes (Pippin et al. 1993). The protein interaction can be detected by using yeast two-hybrid system (Rain et al. 2001).

To Understand a Proteome, Three Distinct Type of Analysis must be Carried Out:

(1) Protein-expression proteomics is the quantitative study of the protein expression of the entire proteome or sub-proteome of two samples that differ by some variable. Identification of novel proteins in signal transduction and disease specific proteins are major outcome of this approach.
(2) Structural proteomics attempts to identify all the proteins within a complex or organelle, determine their localization, and characterize all protein-protein interactions.

The major goal of these studies is to map out the structure of protein complexes or cellular organelle proteins (Blackstock and Weir, 1999).
(3) Functional proteomics allows the study of a selected group of proteins responsible in signalling pathways, diseases and protein-protein interactions. This may be possible by isolating the specific sub-proteomes by affinity-chromatography for further analysis (Fig. 18.2):

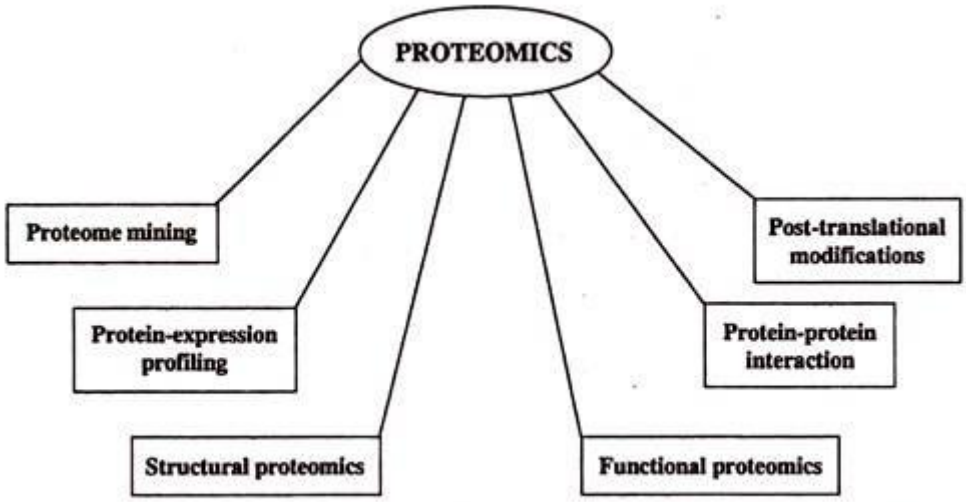


Fig. 18·2 : *Understanding the applications of proteomics (Graves and Haystead, 2002)*

Technology of Proteomics:

Measurement of the level of a gene transcript does not necessarily give clear picture of protein products formed. Therefore, for the measurement of real gene expression, the proteins should be analyzed. Before the identification and measurement of the activity, all the proteins in a proteome for any instant should be separated from each other.

A Typical Proteomics Experiment (e. g. Protein Expression Profiling) can be Divided into the following Categories:

(i) Separation and isolation of protein

(ii) The acquisition of protein structural information for protein identification and characterization

(iii) Database utilization.

(i) Protein Separation and Isolation:

An essential component of proteomics is the protein electrophoresis, the most effective way to resolve a complex mixture of proteins. Two types of electrophoresis are

available as one and two-dimensional electrophoresis. In one dimensional gel electrophoresis (1-DE), proteins are resolved on the basis of their molecular masses. Proteins are stable enough during 1-DE due to their solubility in sodium dodycyl sulphate (SDS). Proteins with molecular mass of 10-300 kDa can be easily separated through 1-DE.

But with complex protein mixtures, results with 1-DE are limited, so for more complex protein mixture such as crude cell lysate, the best separation tool available is two dimensional gel electrophoresis (2-DE) (O'Ferrall, 1975). Here, proteins are separated according to their net charges in first dimension and according to their molecular masses in second dimension.

As a single 2-DE gel can resolve thousands of proteins, it remains a powerful tool for the cataloguing of proteins. Two-dimensional electrophoresis has the ability to resolve proteins that have gone under some post-translational modifications as well as protein expression of any two samples can be compared quantitatively and qualitatively. Recently pH gradients have been introduced to 2-DE which greatly improved the reproducibility of this technique (Bjellqvist et al. 1993).However, few problems with 2-DE still remain to be solved. Despite efforts to automate protein analysis by 2-DE, it is still a labour-intensive and time-consuming process. Another major limitation of 2-DE is the inability to detect low copy number proteins when a total cell lysate is analysed (Link et al. 1997; Shevchenko et al. 1996) as well as inefficiency to speed up the in-gel digestion process also.

Therefore, alternatives have been searched to bypass protein gel electrophoresis. One approach is proteolytic digestion of protein mixture to convert them into peptides and then purify the peptides before subjecting them to analysis by mass spectrometry (MS). Peptide purification has been simplified through liquid chromatography (Link et al. 1999; McCormack et al. 1997), capillary electrophoresis (Figeys et al. 1999; Tong et al. 1999) and reverse phase chromatography (Opiteck et al. 1997).

Recently, Juan et al. (2005) have developed a new approach to speed up the protein identification process utilizing 'microwave' technology. Proteins excised from the gels are subjected to trypsin digestion by microwave irradiation, which rapidly produces peptides fragments. These fragments could be analysed by MALDI (Matrix Assisted Laser Desorption/Ionization). Despite much downstream research on certain alternatives to 2-DE, this is the most widely utilized technique for proteome studies.

(ii) Acquisition of Protein Structures: Protein Identification:

Edman Sequencing (ES):

One of the earliest methods used for protein identification was micro sequencing by Edman chemistry to obtain N-terminal amino acid sequences. This technique was

introduced by Edman in 1949. In Edman sequencing, N-terminal of a protein is sequenced to determine its true start site. Edman sequencing is more applicable sequencing method for the identification of proteins separated by SDS-Polyacrylamide gel electrophoresis.

This method has been used extensively in the starting years of proteomics but certain limitations have emerged in recent time. One of the major limitations is the N-terminal modification of proteins. If any protein is blocked on N-terminal before sequencing, then it is very difficult to identify the protein.

To overcome this problem a novel approach of mixed peptide sequencing (Damer et al. 1998) has been employed recently. In this approach, a protein is converted into peptides by cleavage with cyanogen bromide (CNBr) or skatole followed by the Edman sequencing of peptides.

Mass Spectrometry (MS):

The most significant breakthrough in proteomics has been the mass spectrometric identification of gel-separated proteins. Due to its high sensitivity levels, identification of proteins in protein complexes/mixtures and high throughput, this technique has been proved far better than ES.

In mass spectrometry, proteins are digested into peptides in the gel itself by suitable protease such as trypsin, because proteins, as such, are difficult to elute out from the gels. Moreover, molecular weight of proteins is not usually suitable for database identification. In contrast, peptides can be eluted from the gels easily and matching of even a small set of peptides to the database is quite sufficient to identify a protein.

There are Two Main Approaches to Mass Spectrometric Protein Identification:

(i) “Electrospray ionization” (ESI) involves the fragmentation of individual peptides followed by direct ionization through electrospray in a tandem mass spectrometer. In ESI, a liquid sample flows from a microcapillary tube into the orifice of the mass spectrometer, where a potential difference between the capillary and the inlet to the mass spectrometer results in the generation of a fine mist of charged droplets (Fenn et al. 1989; Hunt et al. 1981).

It has the ability to resolve peptides in a mixture, isolate one species at a time and dissociate it into amino or carboxy-terminal containing fragments designated ‘b’ and ‘y’, respectively.

(ii) In “Peptide mass mapping” approach (Henzel et al. 1993) the mass spectrum of the eluted peptide mixture is acquired, which result in a peptide mass fingerprint of the

protein being studied. The mass spectrum is obtained by a relatively simple 'mass spectrometric method-matrix assisted laser desorption/ ionization' (MALDI).

In this approach, tryptic peptide mixture is analysed because trypsin cleaves proteins at the amino acid arginine and lysine. As the tryptic peptides can be predicted theoretically for any protein, the predicted peptide masses can be compared with those obtained experimentally by MALDI analysis. If the sufficient number of peptide matches with the existing protein sequence in database, the accuracy for protein identification is high.

After the protease cleavages of the proteins, they are analysed by mass analysis also. Mass analysis follows the conversion of proteins or peptides into molecular ions. These ions got separated in a mass spectrometer based on their mass/charge (m/z) ratio. It is determined by the time it takes for the ions to reach the detector. Hence the instrument is called a time of flight (TOF) instrument. The relationship that allows the m/z ratio to be determined is $E = 1/2\ (m/z)v^2$. In this equation. E is the energy imparted on the charged ions as a result of the voltage that is applied by the instrument and V is the velocity of the ions down the flight path. As peptide ions are introduced into the collision chamber, they interact with collision gas and undergo fragmentation along the peptide backbone (Fig. 18.4).

Because all the ions are exposed to the same electric field, all similarly charged ions will have similar energies. Therefore, based on the above equation, ions that have larger mass must have lower velocities and hence will require longer times to reach the detector. Different steps involved in mass spectrometry are described in a flow chart in Fig. 18.3.

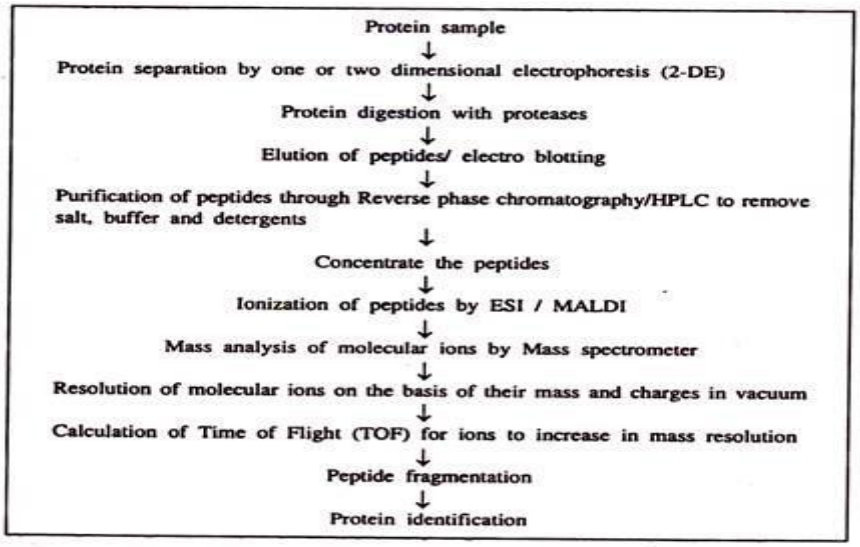


Fig. 18·3 : *A schematic representation of protein identification through Mass spectrometry. All the proteins present in the protein muixture of a cell lysate are identified with this method.*

(iii) Database Utilization:

Initially, sequencing of some proteins or peptides followed by the submission of sequences together created an assembly of proteins called protein database. Proteolytic digestion of many proteins are also predicted theoretically and deposited in database. Hence, at present, so much information has been accumulated that we can search for a homology between a new peptide sequence and the existing sequences in the database to identify the protein.

The major goal of database searching is to identify a large number of proteins—quickly and accurately. All the information accumulated through Edman sequencing or mass spectrometry are used to identify the proteins. In peptide mass fingerprinting database searching, the mass of a unknown peptide after proteolytic digestion is compared to the predicted mass of peptide from theoretical digestion of proteins in database. In amino acid sequence database searching, the sequence of amino acids from a peptide is identified and can be used to search databases to find the protein from which it was derived.

Collection of protein sequence databases are thus designed to represent a partial list of an organism's genome, that is, the genes and all of the proteins they encode. The protein families are usually classified according to their evolutionary history inferred from sequence homology.

These databases are excellent tools for gene discovery, comparative genomics and molecular evolution. The purpose of database similarity searching is the sensitive detection of sequence homologues, regardless of the species relationship in order to infer similarity of function from similarity of sequence.

Recently, Chromatography-based proteomics is used to measure the concentration of low molecular weight peptides in complex mixtures such as plasma or sera. These technologies use time-of-flight (TOF) spectroscopy with matrix-assisted or surface-enhanced laser desorption/ionization to produce a spectrum of mass-to-charge (m/z) ratios that can be analysed in order to identify unique signatures from its chromatography pattern.

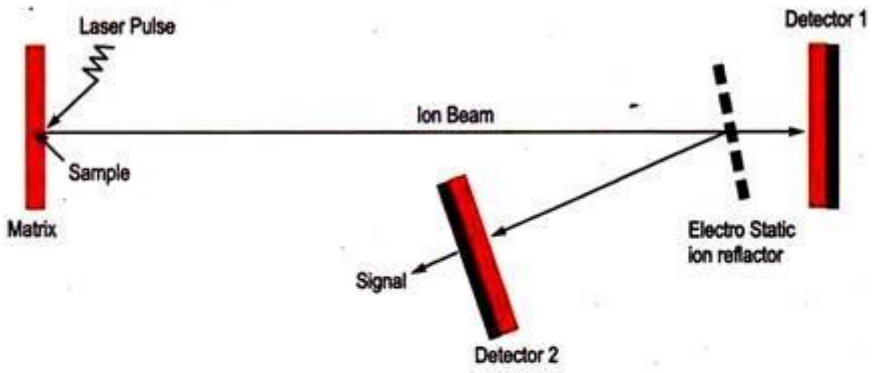


Fig. 18-4 : *Principle behind MALDI-TOF mass spectrometry. A sample is placed on the matrix and ionize by the laser beam. Due to the potential developed between the matrix and the sample, ions start moving towards the detector and get reflected by a reflector in the mid-way. Again after a flight in the tube the ions are detected by another detector. The time taken by these ions in the flight tubes depends on their masses. Therefore, we can calculate the ratio between the mass of an ion and the time of flight in the tube taken by that particular ion*

Applications of Proteomics:

1. Post-Translational Modifications:

Proteomics studies involve certain unique features as the ability to analyze post-translational modifications of proteins. These modifications can be phosphorylation, glycosylation and sulphation as well as some other modifications involved in the maintenance of the structure of a protein.

These modifications are very important for the activity, solubility and localization of proteins in the cell. Determination of protein modification is much more difficult rather than the identification of proteins. As for identification purpose, only few peptides are required for protease cleavages followed by database alignment of a known sequence of a peptide. But for determination of modification in a protein, much more material is needed as all the peptides do not have the expected molecular mass need to be analysed further.

For example, during protein phosphorylation events, phosphopeptides are 80 Da heavier than their unmodified counterparts. Therefore, it gives, rise to a specific fragment (PO^{3-} mass 79) bind to metal resins, get recognized by specific antibodies and later phosphate group can be removed by phosphatases (Clauser et al. 1999; Colledge and Scott, 1999). So protein of interest (post-translationally modified protein) can be detected by Western blotting with the help of antibodies or ^{32}P-labelling that recognize only the active state of molecules. Later, these spots can be identified by mass spectrometry.

2. Protein-Protein Interactions:

The major attribution of proteomics towards the development of protein interactions map of a cell is of immense value to understand the biology of a cell. The knowledge about the time of expression of a particular protein, its level of expression, and, finally, its interaction with another protein to form an intermediate for the performance of a specific biological function is currently available.

These intermediates can be exploited for therapeutic purposes also. An attractive way to study the protein-protein interactions is to purify the entire multi-protein complex by affinity-based methods using GST-fusion proteins, antibodies, peptides etc.

The yeast two-hybrid system has emerged as a powerful tool to study protein-protein interactions (Haynes and Yates, 2000). According to Pandey and Mann (2000) it is a genetic method based on the modular structure of transcription factors in the close proximity of DNA binding domain to the activation domain induces increased transcription of a set of genes.

The yeast hybrid system uses ORFs fused to the DNA binding or activation domain of GAL4 such that increased transcription of a reporter gene results when the proteins encoded by two ORFs interact in the nucleus of the yeast cell. One of the main consequences of this is that once a positive interaction is detected, simply sequencing the relevant clones identifies the ORF. For this reason it is a generic method that is simple and amenable to high throughput screening of protein-protein interactions.Phage display is a method where bacteriophage particles are made to express either a peptide or protein of interest fused to a capsid or coat protein. It can be used to screen for peptide epitopes, peptide ligands, enzyme substrate or single chain antibody fragments.

Another important method to detect protein-protein interactions involves the use of fluorescence resonance energy transfer (FRET) between fluorescent tags on interacting proteins. FRET is a non-radioactive process whereby energy from an excited donor fluorophore is transferred to an acceptor fluorophore. After excitation of the first fluorophore, FRET is detected either by emission from the second fluorophore using appropriate filters or by alteration of the fluorescence lifetime of the donor.

A proteomics strategy of increasing importance involves the localization of proteins in cells as a necessary first step towards understanding protein function in complex cellular networks. The discovery of GFP (green fluorescent protein) and the development of its spectral variants has opened the door to analysis of proteins in living cells by use of the light microscope.Large-scale approaches of localizing GFP-tagged proteins in cells have been performed in the genetically amenable yeast S. pombe (Ding et al. 2000) and in Drosophila (Morin et al. 2001). To localize proteins in mammalian cells, a strategy was developed that enables the systematic GFP tagging of ORFs from novel full-length cDNAs that are identified in genome projects.

3. Protein Expression Profiling:

The largest application of proteomics continues to be protein expression profiling. The expression levels of a protein sample could be measured by 2-DE or other novel technique such as isotope coded affinity tag (ICAT). Using these approaches the varying levels of expression of two different protein samples can also be analysed.

This application of proteomics would be helpful in identifying the signalling mechanisms as well as disease specific proteins. With the help of 2-DE several proteins have been identified that are responsible for heart diseases and cancer (Celis et al. 1999). Proteomics helps in identifying the cancer cells from the non-cancerous cells due to the presence of differentially expressed proteins. The technique of Isotope Coded Affinity Tag has developed new horizons in the field of proteomics. This involves the labelling of two different proteins from two different sources with two chemically identical reagents that differ in their masses due to isotope composition (Gygi et al. 1999). The biggest advantage of this technique is the elimination of protein quantitation by 2-DE. Therefore, high amount of protein sample can be used to enrich low abundance proteins.

Different methods have been used to probe genomic sets of proteins for biochemical activity. One method is called a biochemical genomics approach, which uses parallel biochemical analysis of a proteome comprised of pools of purified proteins in order to identify proteins and the corresponding ORFs responsible for a biochemical activity. The second approach for analysing genomic sets of proteins is the use of functional protein microarrays, in which individually purified proteins are separately spotted on a surface such as a glass slide and then analysed for activity. This approach has huge potential for rapid high-throughput analysis of proteomes and other large collections of proteins, and promises to transform the field of biochemical analysis.

4. Molecular Medicine:

With the help of the information available through clinical proteomics, several drugs have been designed. This aims to discover the proteins with medical relevance to identify a potential target for pharmaceutical development, a marker(s) for disease diagnosis or staging, and risk assessment—both for medical and environmental studies. Proteomic technologies will play an important role in drug discovery, diagnostics and molecular medicine because of the link between genes, proteins and disease.

As researchers study defective proteins that cause particular diseases, their findings will help develop new drugs that either alter the shape of a defective protein or mimic a missing one. Already, many of the best-selling drugs today either act by targeting proteins or are proteins themselves. Advances in proteomics may help scientists eventually create medications that are “personalized” for different individuals to be more effective and have fewer side effects. Current research is looking at protein families linked to disease including cancer, diabetes and heart disease.

Probable Questions:

1. Define DNA Fingerprinting. What are its applications.
2. How DNA Markers are used in disease diagnosis and fingerprinting?
3. Write a note on RFLP with diagram.
4. Write a note on AFLP with diagram.
5. Write a note on RAPD with diagram.
6. Write a note on SNP with diagram.
7. Write a note on VNTR with diagram.
8. Write a note on STR with diagram.
9. What are the types of Proteomics.
10. Write down the significance of proteomics.
11. What you know about Edman Degradation technique?
12. How Mass spectrometry is used in protein study?
13. What are the applications of Proteomics?

Suggested Readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics . Verma and Agarwal.

UNIT-IV

Sociobiology, Altruism, Kin selection and inclusive fitness, Haplodiploidy, Imprinting phenomena

Objective: In this unit we will discuss about Sociobiology, Altruism, Kin selection and inclusive fitness, Haplodiploidy, Imprinting phenomena

Sociobiology:

Sociobiology is based on the premise that some behaviours (social and individual) are at least partly inherited and can be affected by natural selection. It begins with the idea that behaviours have evolved over time, similar to the way that physical traits are thought to have evolved. It predicts that animals will act in ways that have proven to be evolutionarily successful over time. This can, among other things, result in the formation of complex social processes conducive to evolutionary fitness.

The discipline seeks to explain behaviour as a product of natural selection. Behaviour is therefore seen as an effort to preserve one's genes in the population. Inherent in sociobiological reasoning is the idea that certain genes or gene combinations that influence particular behavioural traits can be inherited from generation to generation

For example, newly dominant male lions often kill cubs in the pride that they did not sire. This behaviour is adaptive because killing the cubs eliminates competition for their own offspring and causes the nursing females to come into heat faster, thus allowing more of his genes to enter into the population. Sociobiologists would view this instinctual cub-killing behaviour as being inherited through the genes of successfully reproducing male lions, whereas non-killing behaviour may have died out as those lions were less successful in reproducing. Sociobiologists believe that human behaviour, as well as nonhuman animal behaviour, can be partly explained as the outcome of natural selection. They contend that in order to fully understand behaviour, it must be analysed in terms of evolutionary considerations.

Natural selection is fundamental to evolutionary theory. Variants of hereditary traits which increase an organism's ability to survive and reproduce will be more greatly represented in subsequent generations, i.e., they will be "selected for". Thus, inherited behavioural mechanisms that allowed an organism a greater chance of surviving and/or reproducing in the past are more likely to survive in present organisms. That inherited adaptive behaviours are present in nonhuman animal species has been multiply demonstrated by biologists, and it has become a foundation of evolutionary biology. However, there is continued resistance by some researchers over the application of evolutionary models to humans, particularly from within the social sciences, where culture has long been assumed to be the predominant driver of behaviour.

Sociobiology is based upon two fundamental premises:

- Certain behavioural traits are inherited,
- Inherited behavioural traits have been honed by natural selection. Therefore, these traits were probably "adaptive" in the environment in which the species evolved.

Sociobiology uses Nikolaas Tinbergen's four categories of questions and explanations of animal behaviour. Two categories are at the species level; two, at the individual level. The species-level categories (often called "ultimate explanations") are

- the function (i.e., adaptation) that a behaviour serves and
- the evolutionary process (i.e., phylogeny) that resulted in this functionality. The individual-level categories (often called "proximate explanations") are
- the development of the individual (i.e., ontogeny) and
- the proximate mechanism (e.g., brain anatomy and hormones).

Sociobiologists are interested in how behaviour can be explained logically as a result of selective pressures in the history of a species. Thus, they are often interested in instinctive, or intuitive behaviour, and in explaining the similarities, rather than the differences, between cultures. For example, mothers within many species of mammals – including humans – are very protective of their offspring. Sociobiologists reason that this protective behaviour likely evolved over time because it helped the offspring of the individuals which had the characteristic to survive. This parental protection would increase in frequency in the population. The social behaviour is believed to have evolved in a fashion similar to other types of nonbehavioural adaptations, such as a coat of fur, or the sense of smell.

Individual genetic advantage fails to explain certain social behaviours as a result of gene-centred selection. E.O. Wilson argued that evolution may also act upon groups The mechanisms responsible for group selection employ paradigms and population statistics borrowed from evolutionary game theory. Altruism is defined as "a concern for the welfare of others". If altruism is genetically determined, then altruistic individuals must reproduce their own altruistic genetic traits for altruism to survive, but when altruists lavish their resources on non-altruists at the expense of their own kind, the altruists tend to die out and the others tend to increase. An extreme example is a soldier losing his life trying to help a fellow soldier. This example raises the question of how altruistic genes can be passed on if this soldier dies without having any children.

Within sociobiology, a social behaviour is first explained as a sociobiological hypothesis by finding an evolutionarily stable strategy that matches the observed behaviour. Stability of a strategy can be difficult to prove, but usually, it will predict gene frequencies. The hypothesis can be supported by establishing a correlation

between the gene frequencies predicted by the strategy, and those expressed in a population. Altruism between social insects and littermates has been explained in such a way. Altruistic behaviour, behaviour that increases the reproductive fitness of others at the apparent expense of the altruist, in some animals has been correlated to the degree of genome shared between altruistic individuals. A quantitative description of infanticide by male harem-mating animals when the alpha male is displaced as well as rodent female infanticide and fetal resorption are active areas of study. In general, females with more bearing opportunities may value offspring less, and may also arrange bearing opportunities to maximize the food and protection from mates. An important concept in sociobiology is that temperament traits exist in an ecological balance. Just as an expansion of a sheep population might encourage the expansion of a wolf population, an expansion of altruistic traits within a gene pool may also encourage increasing numbers of individuals with dependent traits.

Studies of human behaviour genetics have generally found behavioural traits such as creativity, extroversion, aggressiveness, and IQ have high heritability. The researchers who carry out those studies are careful to point out that heritability does not constrain the influence that environmental or cultural factors may have on those traits.

Altruism:

Altruism is the principle and moral practice of concern for happiness of other human beings, resulting in a quality of life both material and spiritual. It is a traditional virtue in many cultures and a core aspect of various religious traditions and secular worldviews, though the concept of "others" toward whom concern should be directed can vary among cultures and religions. In an extreme case, altruism may become a synonym of selflessness which is the opposite of selfishness.
In a common way of living, it doesn't deny the singular nature of the subject, but realizes the traits of the individual personality in relation to the others, with a true, direct and personal interaction with each of them. It is focusing both on the single people and the whole community. In an (not only) Christian practice, it is the law of love direct to the ego and his neighbour. The word "altruism" was coined by the French philosopher Auguste Comte in French, as *altruisme*, for an antonym of egoism. He derived it from the Italian *altrui*, which in turn was derived from Latin *alteri*, meaning "other people" or "somebody else".
Altruism in biological observations in field populations of the day organisms can be defined as an individual performing an action which is at a cost to themselves (e.g., pleasure and quality of life, time, probability of survival or reproduction), but benefits, either directly or indirectly, another third-party individual, without the expectation of reciprocity or compensation for that action. Steinberg suggests a definition for altruism in the clinical setting, that is "intentional and voluntary actions that aim to enhance the welfare of another person in the absence of any quid pro quo external rewards". Altruism can be distinguished from feelings of loyalty, in that whilst

the latter is predicated upon social relationships, altruism does not consider relationships. Much debate exists as to whether *"true"* altruism is possible in human psychology. The theory of psychological egoism suggests that no act of sharing, helping or sacrificing can be described as truly altruistic, as the actor may receive an intrinsic reward in the form of personal gratification. The validity of this argument depends on whether intrinsic rewards qualify as "benefits".
The term *altruism* may also refer to an ethical doctrine that claims that individuals are morally obliged to benefit others. Used in this sense, it is usually contrasted with egoism, which claims individuals are morally obligated to serve themselves first. The concept has a long history in philosophical and ethical thought. The term was originally coined in the 19th century by the founding sociologist and philosopher of science, Auguste Comte, and has become a major topic for psychologists (especially evolutionary psychology researchers), evolutionary biologists, and ethologists. Whilst ideas about altruism from one field can affect the other fields, the different methods and focuses of these fields always lead to different perspectives on altruism. In simple terms, altruism is caring about the welfare of other people and acting to help them.

Kin Selection:

Kin selection is the evolutionary strategy that favours the reproductive success of an organism's relatives, even at a cost to the organism's own survival and reproduction. Kin altruism can look like altruistic behaviour whose evolution is driven by kin selection. Kin selection is an instance of inclusive fitness, which combines the number of offspring produced with the number an individual can ensure the production of by supporting others, such as siblings. John Maynard Smith may have coined the actual term "kin selection" in 1964.
Charles Darwin discussed the concept of kin selection in his 1859 book, *The Origin of Species*, where he reflected on the puzzle of sterile social insects, such as honey bees, which leave reproduction to their mothers, arguing that a selection benefit to related organisms (the same "stock") would allow the evolution of a trait that confers the benefit but destroys an individual at the same time. R.A. Fisher in 1930 and J.B.S. Haldane in 1932 set out the mathematics of kin selection, with Haldane famously joking that he would willingly die for two brothers or eight cousins. In 1964, W.D. Hamilton popularised the concept and the major advance in the mathematical treatment of the phenomenon by George R. Price which has become known as Hamilton's rule. In the same year John Maynard Smith used the actual term kin selection for the first time.

According to Hamilton's rule, kin selection causes genes to increase in frequency when the genetic relatedness of a recipient to an actor multiplied by the benefit to the recipient is greater than the reproductive cost to the actor. Hamilton proposed two mechanisms for kin selection. First, kin recognition allows individuals to be able to identify their relatives. Second, in viscous populations, populations in which the movement of organisms from their place of birth is relatively slow, local interactions

tend to be among relatives by default. The viscous population mechanism makes kin selection and social cooperation possible in the absence of kin recognition. In this case, nurture kinship, the treatment of individuals as kin as a result of living together, is sufficient for kin selection, given reasonable assumptions about population dispersal rates. Note that kin selection is not the same thing as group selection, where natural selection is believed to act on the group as a whole.
In humans, altruism is both more likely and on a larger scale with kin than with unrelated individuals; for example, humans give presents according to how closely related they are to the recipient. In other species, vervet monkeys use allomothering, where related females such as older sisters or grandmothers often care for young, according to their relatedness. The social shrimp *Synalpheusregalis* protects juveniles within highly related colonies.

The earliest mathematically formal treatments of kin selection were by R.A. Fisher in 1930 and J.B.S. Haldane in 1932and 1955.J.B.S. Haldane fully grasped the basic quantities and considerations in kin selection, famously writing "I would lay down my life for two brothers or eight cousins". Haldane's remark alluded to the fact that if an individual loses its life to save two siblings, four nephews, or eight cousins, it is a "fair deal" in evolutionary terms, as siblings are on average 50% identical by descent, nephews 25%, and cousins 12.5% (in a diploid population that is randomly mating and previously outbred). But Haldane also joked that he would truly die only to save more than a single identical twin of his or more than two full siblings. In 1955 he clarified:

Let us suppose that you carry a rare gene that affects your behaviour so that you jump into a flooded river and save a child, but you have one chance in ten of being drowned, while I do not possess the gene, and stand on the bank and watch the child drown. If the child's your own child or your brother or sister, there is an even chance that this child will also have this gene, so five genes will be saved in children for one lost in an adult. If you save a grandchild or a nephew, the advantage is only two and a half to one. If you only save a first cousin, the effect is very slight. If you try to save your first cousin once removed the population is more likely to lose this valuable gene than to gain it. ... It is clear that genes making for conduct of this kind would only have a chance of spreading in rather small populations when most of the children were fairly near relatives of the man who risked his life.

Inclusive fitness:

In evolutionary biology, inclusive fitness is one of two metrics of evolutionary success as defined by W. D. Hamilton in 1964:

- **Personal fitness** is the number of offspring that an individual begets (regardless of who rescues/rears/supports them)
- **Inclusive fitness** is the number of offspring equivalents that an individual rears, rescues or otherwise supports through its behaviour (regardless of who begets them)

An individual's own child, who carries one half of the individual's genes, is defined

as one offspring equivalent. A sibling's child, who will carry one-quarter of the individual's genes, is 1/2 offspring equivalent. Similarly, a cousin's child, who has 1/16 of the individual's genes, is 1/8 offspring equivalent.

From the gene's point of view, evolutionary success ultimately depends on leaving behind the maximum number of copies of itself in the population. Prior to Hamilton's work, it was generally assumed that genes only achieved this through the number of viable offspring produced by the individual organism they occupied. However, this overlooked a wider consideration of a gene's success, most clearly in the case of the social insects where the vast majority of individuals do not produce offspring.

Hamilton showed mathematically that, because other members of a population may share one's genes, a gene can also increase its evolutionary success by indirectly promoting the reproduction and survival of other individuals who also carry that gene. This is variously called "kin theory", "kin selection theory" or "inclusive fitness theory". The most obvious category of such individuals is close genetic relatives, and where these are concerned, the application of inclusive fitness theory is often more straightforwardly treated via the narrower kin selection theory.
Hamilton's theory, alongside reciprocal altruism, is considered one of the two primary mechanisms for the evolution of social behaviours in natural species and a major contribution to the field of sociobiology, which holds that some behaviours can be dictated by genes, and therefore can be passed to future generations and may be selected for as the organism evolves.

Although described in seemingly anthropomorphic terms, these ideas apply to all living things, and can describe the evolution of innate and learned behaviours over a wide range of species including insects, small mammals or humans.

Belding's ground squirrel provides an example. The ground squirrel gives an alarm call to warn its local group of the presence of a predator. By emitting the alarm, it gives its own location away, putting itself in more danger. In the process, however, the squirrel may protect its relatives within the local group (along with the rest of the group). Therefore, if the effect of the trait influencing the alarm call typically protects the other squirrels in the immediate area, it will lead to the passing on of more copies of the alarm call trait in the next generation than the squirrel could leave by reproducing on its own. In such a case natural selection will increase the trait that influences giving the alarm call, provided that a sufficient fraction of the shared genes include the gene(s) predisposing to the alarm call. *Synalpheus regalis*, a eusocial shrimp, also is an example of an organism whose social traits meet the inclusive fitness criterion. The larger defenders protect the young juveniles in the colony from outsiders. By ensuring the young's survival, the genes will continue to be passed on to future generations. Inclusive fitness is more generalized than strict kin selection, which requires that the shared genes are *identical by descent*. Inclusive fitness is not limited to cases where "kin" ('close genetic relatives') are involved.

Haplodiploidy:

Approximately 15% of all arthropods reproduce through haplodiploidy. Yet it is unclear how this mode of reproduction affects other aspects of reproductive ecology. In this review we outline predictions on how haplodiploidy might affect mating system evolution, the evolution of traits under sexual or sexual antagonistic selection, sex allocation decisions and the evolution of parental care. We also give an overview of the phylogenetic distribution of haplodiploidy. Finally, we discuss how comparisons between different types of haplodiploidy (arrhenotoky, PGE with haploid vs somatically diploid males) might help to discriminate between the effects of virgin birth, haploid gene expression and those of haploid gene transmission.

Haplodiploidy is a sex-determination system in which males develop from unfertilized eggs and are haploid, and females develop from fertilized eggs and are diploid. Haplodiploidy is sometimes called arrhenotoky. Haplodiploidy determines the sex in all members of the insect orders Hymenoptera (bees, ants, and wasps) and Thysanoptera ('thrips'). The system also occurs sporadically in some spider mites, Hemiptera, Coleoptera (bark beetles), and rotifers.
In this system, sex is determined by the number of sets of chromosomes an individual receives. An offspring formed from the union of a sperm and an egg develops as a female, and an unfertilized egg develops as a male. This means that the males have half the number of chromosomes that a female has, and are haploid. The haplodiploid sex-determination system has a number of peculiarities. For example, a male has no father and cannot have sons, but he has a grandfather and can have grandsons. Additionally, if a eusocial-insect colony has only one queen, and she has only mated once, then the relatedness between workers (diploid females) in a hive or nest is 3/4. This means the workers in such monogamous single-queen colonies are significantly more closely related than in other sex determination systems where the relatedness of siblings is usually no more than 1/2. It is this point which drives the kin selection theory of how eusociality evolved. Whether haplodiploidy did in fact pave the way for the evolution of eusociality is still a matter of debate.

Another feature of the haplodiploidy system is that recessive lethal and deleterious alleles will be removed from the population rapidly because they will automatically be expressed in the males (dominant lethal and deleterious alleles are removed from the population every time they arise, as they kill any individual they arise in). Haplodiploidy is not the same thing as an X0 sex- determination system. In haplodiploidy, males receive one half of the chromosomes that females receive, including autosomes. In an X0 sex-determination system, males and females receive an equal number of autosomes, but when it comes to sex chromosomes, females will receive two X chromosomes while males will receive only a single X chromosome.

Several models have been proposed for the genetic mechanisms of haplodiploid sex- determination. The model most commonly referred to is the

complementary allele model. According to this model, if an individual is heterozygous for a certain locus, it develops into a female, whereas hemizygous and homozygous individuals develop into males. In other words, diploid offspring develop from fertilized eggs, and are normally female, while haploid offspring develop into males from unfertilized eggs. Diploid males would be infertile, as their cells would not undergo meiosis to form sperm. Therefore, the sperm would be diploid, which means that their offspring would be triploid. Since hymenopteran mother and sons share the same genes, they may be especially sensitive to inbreeding: Inbreeding reduces the number of different sex alleles present in a population, hence increasing the occurrence of diploid males.

After mating, each fertile hymenopteran female stores sperm in an internal sac called the spermatheca. The mated female controls the release of stored sperm from within the organ: If she releases sperm as an egg passes down her oviduct, the egg is fertilized. Social bees, wasps, and ants can modify sex ratios within colonies which maximizes relatedness among members and generates a workforce appropriate to surrounding conditions. In other solitary hymenopterans, the females lay unfertilized male eggs on poorer food sources while laying the fertilized female eggs on better food sources, possibly because the fitness of females will be more adversely affected by shortages in their early life. Sex ratio manipulation is also practiced by haplodiploid ambrosia beetles, who lay more male eggs when the chances for males to disperse and mate with females in different sites are greater.

Imprinting Phenomena:

Genomic imprinting is an epigenetic phenomenon that causes genes to be expressed in a parent- of-origin-specific manner. Forms of genomic imprinting have been demonstrated in fungi, plants and animals. As of 2014, there are about 150 imprinted genes known in the mouse and about half that in humans.

Genomic imprinting is an inheritance process independent of the classical Mendelian inheritance. It is an epigenetic process that involves DNA methylation and histone methylation without altering the genetic sequence. These epigenetic marks are established ("imprinted") in the germline (sperm or egg cells) of the parents and are maintained through mitotic cell divisions in the somatic cells of an organism. Appropriate imprinting of certain genes is important for normal development. Human diseases involving genomic imprinting include Angelman syndrome and Prader–Willi syndrome.

Imprinting mechanisms:

Imprinting is a dynamic process. It must be possible to erase and re-establish imprints through each generation so that genes that are imprinted in an adult may still be expressed in that adult's offspring. (For example, the maternal genes that control insulin production will be imprinted in a male but will be expressed in any of the male's offspring that inherit these genes.) The nature of imprinting must therefore be epigenetic

rather than DNA sequence dependent. In germline cells the imprint is erased and then re-established according to the sex of the individual, i.e. in the developing sperm (during spermatogenesis), a paternal imprint is established, whereas in developing oocytes (oogenesis), a maternal imprint is established. This process of erasure and reprogramming is necessary such that the germ cell imprinting status is relevant to the sex of the individual. In both plants and mammals there are two major mechanisms that are involved in establishing the imprint; these are DNA methylation and histone modifications.
Recently, a new study has suggested a novel inheritable imprinting mechanism in humans that would be specific of placental tissue and that is independent of DNA methylation (the main and classical mechanism for genomic imprinting). Among the hypothetical explanations for this exclusively human phenomenon, two possible mechanisms have been proposed: either a histone modification that confers imprinting at novel placental-specific imprinted *loci* or, alternatively, a recruitment of DNMTs to these loci by a specific and unknown transcription factor that would be expressed during early trophoblast differentiation.

Hypotheses on the origins of imprinting

A widely accepted hypothesis for the evolution of genomic imprinting is the "parental conflict hypothesis". Also known as the kinship theory of genomic imprinting, this hypothesis states that the inequality between parental genomes due to imprinting is a result of the differing interests of each parent in terms of the evolutionary fitness of their genes The father's genes that encode for imprinting gain greater fitness through the success of the offspring, at the expense of the mother. The mother's evolutionary imperative is often to conserve resources for her own survival while providing sufficient nourishment to current and subsequent litters. Accordingly, paternally expressed genes tend to be growth-promoting whereas maternally expressed genes tend to be growth-limiting. In support of this hypothesis, genomic imprinting has been found in all placental mammals, where post-fertilisation offspring resource consumption at the expense of the mother is high; although it has also been found in oviparous birds where there is relatively little post-fertilisation resource transfer and therefore less parental conflict.

However, our understanding of the molecular mechanisms behind genomic imprinting show that it is the maternal genome that controls much of the imprinting of both its own and the paternally- derived genes in the zygote, making it difficult to explain why the maternal genes would willingly relinquish their dominance to that of the paternally-derived genes in light of the conflict hypothesis. Another hypothesis proposed is that some imprinted genes act co-adaptively to improve both foetal development and maternal provisioning for nutrition and care. In it a subset of paternally expressed genes are co-expressed in both the placenta and the mother's hypothalamus. This would come about through selective pressure from parent-infant coadaptation to

improve infant survival. Paternally expressed 3 (Peg3) is a gene for which this hypothesis may apply.

Others have approached their study of the origins of genomic imprinting from a different side, arguing that natural selection is operating on the role of epigenetic marks as machinery for homologous chromosome recognition during meiosis, rather than on their role in differential expression. This argument centers on the existence of epigenetic effects on chromosomes that do not directly affect gene expression, but do depend on which parent the chromosome originated from. This group of epigenetic changes that depend on the chromosome's parent of origin (including both those that affect gene expression and those that do not) are called parental origin effects, and include phenomena such as paternal X inactivation in the marsupials, nonrandom parental chromatid distribution in the ferns, and even mating type switching in yeast. This diversity in organisms that show parental origin effects has prompted theorists to place the evolutionary origin of genomic imprinting before the last common ancestor of plants and animals, over a billion years ago.

Natural selection for genomic imprinting requires genetic variation in a population. A hypothesis for the origin of this genetic variation states that the host-defense system responsible for silencing foreign DNA elements, such as genes of viral origin, mistakenly silenced genes whose silencing turned out to be beneficial for the organism. There appears to be an over-representation of retrotransposed genes, that is to say genes that are inserted into the genome by viruses, among imprinted genes. It has also been postulated that if the retrotransposed gene is inserted close to another imprinted gene, it may just acquire this imprint

Imprinted genes in Mammals :

That imprinting might be a feature of mammalian development was suggested in breeding experiments in mice carrying reciprocal chromosomal translocations. Nucleus transplantation experiments in mouse zygotes in the early 1980s confirmed that normal development requires the contribution of both the maternal and paternal genomes. The vast majority of mouse embryos derived from parthenogenesis (called parthenogenones, with two maternal or egg genomes) and androgenesis (called androgenones, with two paternal or sperm genomes) die at or before the blastocyst/implantation stage. In the rare instances that they develop to post-implantation stages, gynogenetic embryos show better embryonic development relative to placental development, while for androgenones, the reverse is true. Nevertheless, for the latter, only a few have been described.
No naturally occurring cases of parthenogenesis exist in mammals because of imprinted genes. However, in 2004, experimental manipulation by Japanese researchers of a paternal methylation imprint controlling the *Igf2* gene led to the birth of a mouse (named Kaguya) with two maternal sets of chromosomes, though it is not a true parthenogenone since cells from two different female mice were used. The

researchers were able to succeed by using one egg from an immature parent, thus reducing maternal imprinting, and modifying it to express the gene Igf2, which is normally only expressed by the paternal copy of the gene.Parthenogenetic/gynogenetic embryos have twice the normal expression level of maternally derived genes, and lack expression of paternally expressed genes, while the reverse is true for androgenetic embryos. It is now known that there are at least 80 imprinted genes in humans and mice, many of which are involved in embryonic and placental growth and development. Hybrid offspring of two species may exhibit unusual growth due to the novel combination of imprinted genes.

Various methods have been used to identify imprinted genes. In swine, Bischoff *et al.* 2009 compared transcriptional profiles using short-oligonucleotide microarrays to survey differentially expressed genes between parthenotes (2 maternal genomes) and control fetuses (1 maternal, 1 paternal genome). An intriguing study surveying the transcriptome of murine brain tissues revealed over 1300 imprinted gene loci (approximately 10-fold more than previously reported) by RNA-sequencing from F1 hybrids resulting from reciprocal crosses. The result however has been challenged by others who claimed that this is an overestimation by an order of magnitude due to flawed statistical analysis. In domesticated livestock, single-nucleotide polymorphisms in imprinted genes influencing foetal growth and development have been shown to be associated with economically important production traits in cattle, sheep and pigs.

Probable Questions:

1. Write a note about altruism. What is its significance.
2. What do you mean by kin selection? what is its significance.
3. What is Haplodiploidy?
4. What is Genetic imprinting? How it affects human behaviour?
5. Write about hypothesis on genetic imprinting.
6. Define and explain sociobiology.
7. Write a note on inclusive fitness.

Suggested readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics . Verma and Agarwal.

UNIT-V

Molecular and biochemical basis of genetic diseases: Autosomal (cystic fibrosis), X-linked (haemophilia A), Metabolic disorders (phenylketonuria).

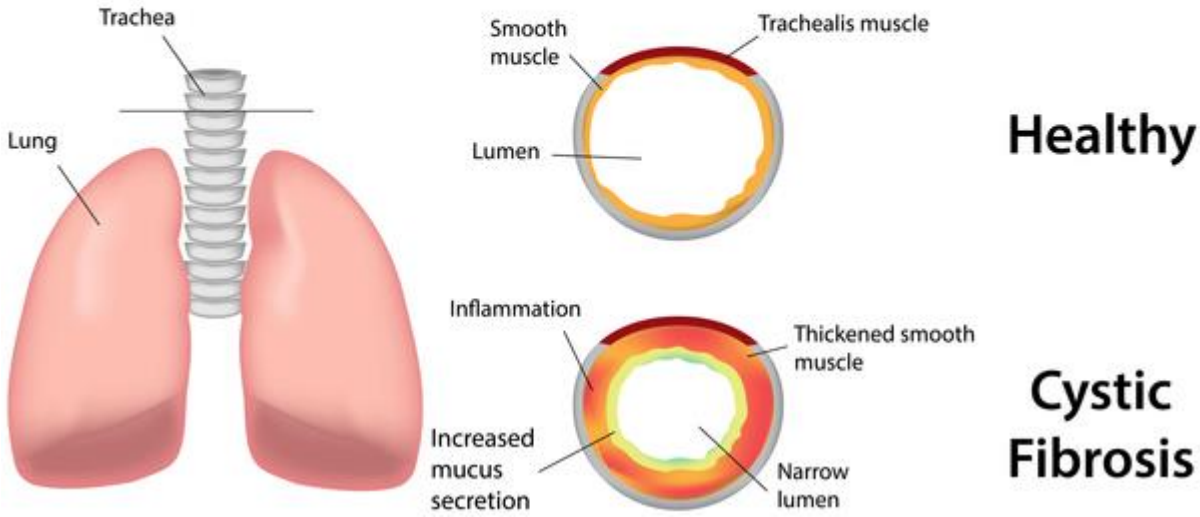
Objective: In this unit we will discuss about Molecular and biochemical basis of some genetic diseases such as cystic fibrosis, haemophilia A and phenylketonuria.

Cystic Fibrosis:

Cystic fibrosis is an inherited disease characterized by the buildup of thick, sticky mucus that can damage many of the body's organs. The disorder's most common signs and symptoms include progressive damage to the respiratory system and chronic digestive system problems. The features of the disorder and their severity vary among affected individuals.

Mucus is a slippery substance that lubricates and protects the linings of the airways, digestive system, reproductive system, and other organs and tissues. In people with cystic fibrosis, the body produces mucus that is abnormally thick and sticky. This abnormal mucus can clog the airways, leading to severe problems with breathing and bacterial infections in the lungs. These infections cause chronic coughing, wheezing, and inflammation. Over time, mucus buildup and infections result in permanent lung damage, including the formation of scar tissue (fibrosis) and cysts in the lungs.

Cystic Fibrosis



Most people with cystic fibrosis also have digestive problems. Some affected babies have meconium ileus, a blockage of the intestine that occurs shortly after birth. Other digestive problems result from a buildup of thick, sticky mucus in the pancreas. The pancreas is an organ that produces insulin (a hormone that helps control blood sugar levels). It also makes enzymes that help digest food. In people with cystic fibrosis, mucus often damages the pancreas, impairing its ability to produce insulin and digestive enzymes. Problems with digestion can lead to diarrhea, malnutrition, poor growth, and weight loss. In adolescence or adulthood, a shortage of insulin can cause a form of diabetes known as cystic fibrosis-related diabetes mellitus (CFRDM).

Cystic fibrosis used to be considered a fatal disease of childhood. With improved treatments and better ways to manage the disease, many people with cystic fibrosis now live well into adulthood. Adults with cystic fibrosis experience health problems affecting the respiratory, digestive, and reproductive systems. Most men with cystic fibrosis have congenital bilateral absence of the vas deferens (CBAVD), a condition in which the tubes that carry sperm (the vas deferens) are blocked by mucus and do not develop properly. Men with CBAVD are unable to father children (infertile) unless they undergo fertility treatment. Women with cystic fibrosis may experience complications in pregnancy.

Frequency: Cystic fibrosis is a common genetic disease within the white population in the United States. The disease occurs in 1 in 2,500 to 3,500 white newborns. Cystic

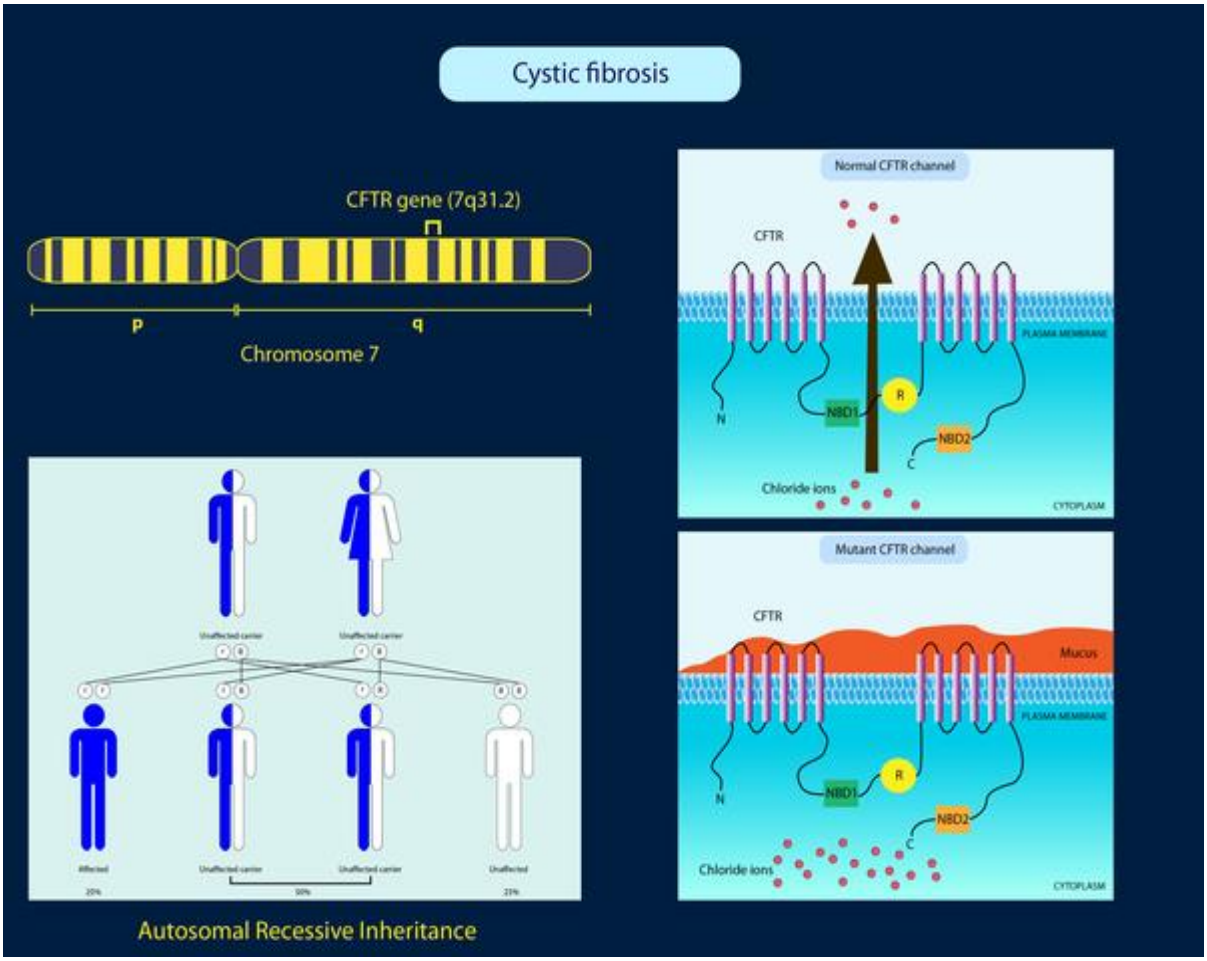
fibrosis is less common in other ethnic groups, affecting about 1 in 17,000 African Americans and 1 in 31,000 Asian Americans.

Characteristics of Cystic Fibrosis:

- Cystic fibrosis (CF) is one of the genetic diseases i.e. it can be inherited to offspring.
- It is the condition where the mucus produced is unusually thick and sticky that mainly affects the lungs and digestive systems along with other body organs.
- CF affects the exocrine glands such as-sweat glands, mucus secreting glands and digestive juice secreting cells.
- Normally, the mucus produced is thin and slippery but in case of CF, the defective gene leads to the creation of thick and sticky mucus.
- The so formed thick and sticky mucus blocks the respiratory ducts resulting difficult breathing and also interferes with digestive function of pancreas.
- Not only lungs and pancreas, people with CF may also face male infertility as the thick mucus causes the blockage of the Vas deferens, or epididymis preventing release of sperm from testis.
- The term Cystic fibrosis was given as the thick mucus built up inside respiratory tract that cause severe lung damage-cysts formation and fibrosis.
- CF is known to appear in almost all ethnic communities but is most common among Caucasians of Northern European descent.

Causes:

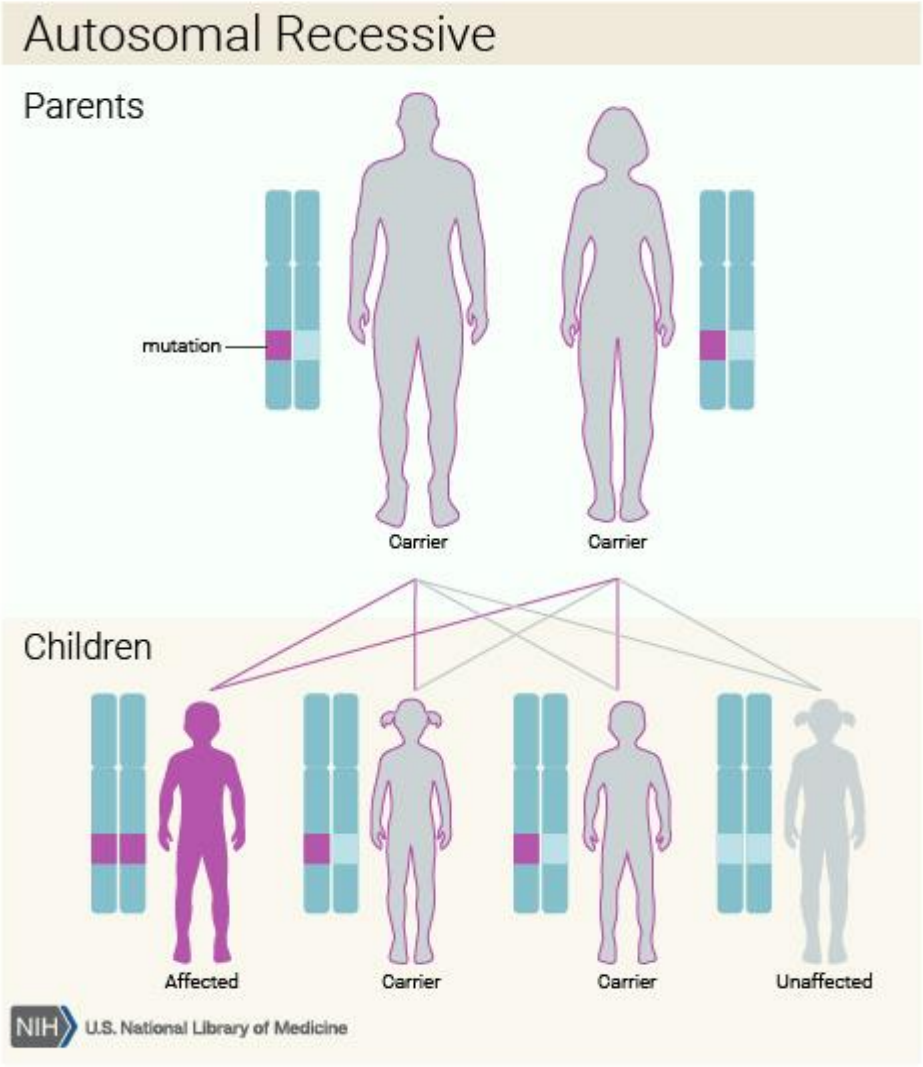
Mutations in the *CFTR* gene cause cystic fibrosis. The *CFTR* gene provides instructions for making a channel that transports negatively charged particles called chloride ions into and out of cells.



Chloride is a component of sodium chloride, a common salt found in sweat. Chloride also has important functions in cells; for example, the flow of chloride ions helps control the movement of water in tissues, which is necessary for the production of thin, freely flowing mucus. Mutations in the *CFTR* gene disrupt the function of the chloride channels, preventing them from regulating the flow of chloride ions and water across cell membranes. As a result, cells that line the passageways of the lungs, pancreas, and other organs produce mucus that is unusually thick and sticky. This mucus clogs the airways and various ducts, causing the characteristic signs and symptoms of cystic fibrosis. Other genetic and environmental factors likely influence the severity of the condition. For example, mutations in genes other than *CFTR* might help explain why some people with cystic fibrosis are more severely affected than others. Most of these genetic changes have not been identified, however.

Inheritance:

This condition is inherited in an autosomal recessive pattern, which means both copies of the gene in each cell have mutations. The parents of an individual with an autosomal recessive condition each carry one copy of the mutated gene, but they typically do not show signs and symptoms of the condition.



Symptoms:

On the basis of severity of disorder, symptoms vary from individual to individual. Here, the classification is done as respiratory and digestive symptoms.

a. Respiratory symptoms:

Lung infections

Stuffy nose due to inflammation of nasal passages.

Frequent occurrence of sinusitis

Fatigue

Persistent coughing along with thick mucus

Bronchitis

Difficulty in breathing/ Wheezing

Pneumonia

b. Digestive Symptoms:

Diarrhea or foul odor and greasy stools

Gastritis

Constipation

Inflammation of pancreas

Nausea

Loss of appetite

Slow growth and development in child

Weight loss causing Malnutrition

Bowel obstruction in new born babies

c. Other Symptoms:

Salty test of skin due to salty sweat

Infertility in male

Osteoporosis

Liver problems

Diabetes

Difficulty during pregnancy

Diagnosis:

Cystic fibrosis can be diagnosed by the help of various tests:

a. Immunoreactive trypsinogen test (IRT)

IRT is a screening test performed for new born babies to detect the levels of protein called IRT in the blood.

It is the standard test for newborn screening. The high levels of IRT denotes cystic fibrosis. However, further testing should be proceeded for confirmation of CF.

b. Sweat chloride test:

The elevated levels of chloride in the sweat is diagnosis of CF.

Sweat chloride test is executed by using a chemical that produces sweat when triggered by relatively weak electric current.

Sweat is accumulated either in paper or pad and then analyzed.

Saltier sweat than normal is diagnosis for CF.

c. Chest X-ray:

Blockade of respiratory tracts causes the inflammation of lungs and this can be disclosed by chest X-rays.

d. Sputum test:

A sample of mucus is taken and it is examined to check whether it is infected or not.

This test also shows the microorganisms present and determines the antibiotics for treatment.

e. CT Scan:

CT scan displays the detailed view of the internal structures such as pancreas and lungs.

This helps to assess the range of damage caused by cystic fibrosis.

f. Pulmonary Function tests:

The function of lungs such as inhale and exhale of air and the transportation of oxygen to various parts is analyzed by these tests.

Any unusual response may indicate cystic fibrosis.

g. Genetic tests:

This test is carried out by checking a sample of blood or cheek cells which checks for defective gene causing cystic fibrosis.

This test also reveals carrier of cystic fibrosis.

Treatment:

There is no any known cure for cystic fibrosis till date. However, treatments are prevalent for the control and management of symptoms in order to provide a quality life for individuals with CF. Implanted devices are used for long term administration of drugs. CF transmembrane conductance regulator (CFTR) modulators are latest medications to target the defective CFTR gene. Aerobic exercise is advised that involves harder breathing for the release of mucus from airways.

a. Air passage clearance:

It is very difficult to get rid of mucus for individuals with CF and they have hard time breathing. In order to clear the mucus and allow them clear breathing, airway clearance techniques (ACT) are used. ACT also minimizes lung infection. For instance, a therapist strokes the chest and help to free mucus. Also oral medications are prescribed to thin the mucus, mobilize it and kill germs. Azithromycin and ibuprofen are prescribed for enhancing lung function. Bronchodilators help to clear the airways by relaxing muscles around them.

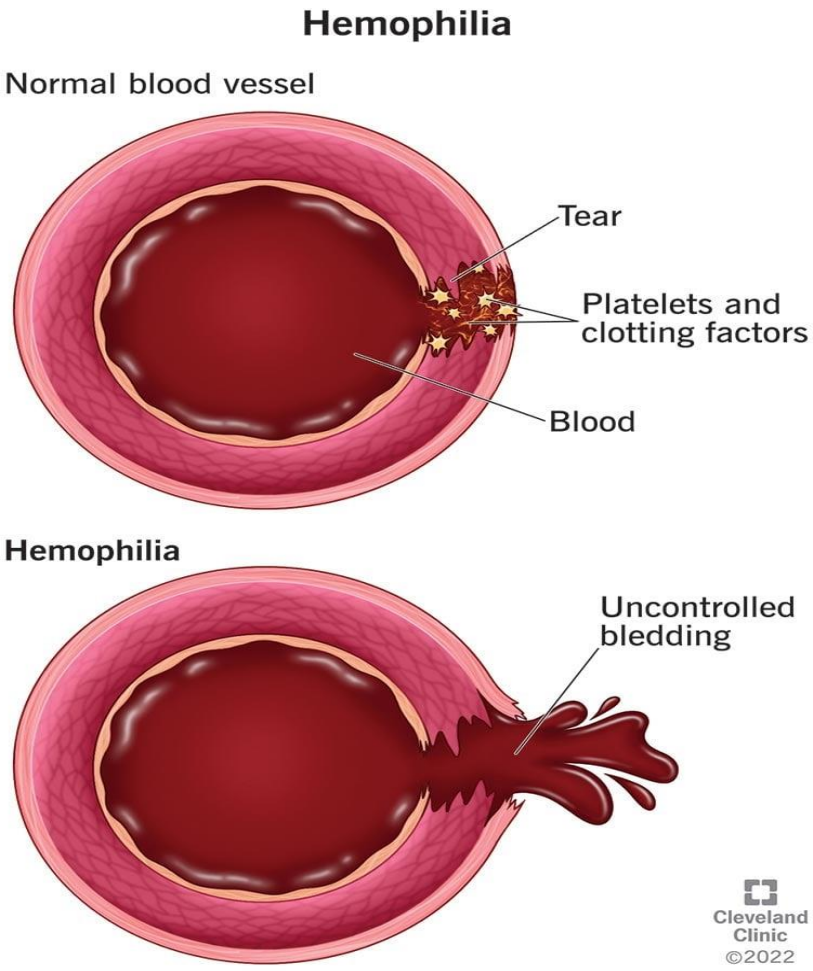
b. Nutritional therapy:

CF is responsible for the disturbance in digestive system and absorption of nutrients. Individuals with CF need to discuss about proper diet with the dietician. Pancreatic or other digestive supplements if needed should be balanced for proper digestive absorption. High calorie and high fat diet is recommended for children with CF for their proper growth and development.

Haemophilia:

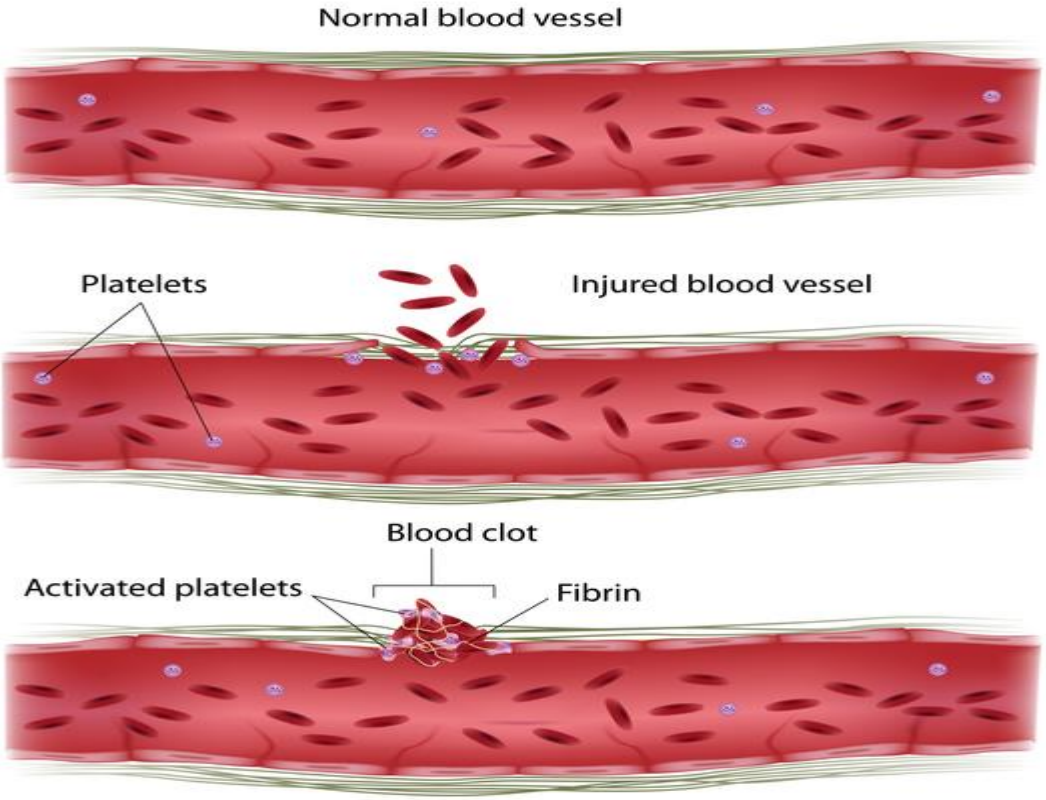
It is commonly known as bleeder's disease. It is a sex linked disease first studied by John Cotto in 1803. Generally male individuals are the victim of this disease. Due to the absence of antihaemophiliac globulin (in case of haemophilia-A) or plasma thromboplastin (haemophilia-B) in the blood the patient bleeds for hours (in normal man it takes 2-8 minutes to clot) even from a minor cut. As a result of continuous bleeding, the patient may die of blood loss. Although there is no permanent remedy of this disease, transfusion of normal blood helps in checking bleeding by providing required blood clotting factors. For patients suffering from haemophilia A antihaemophiliac globulin is available.

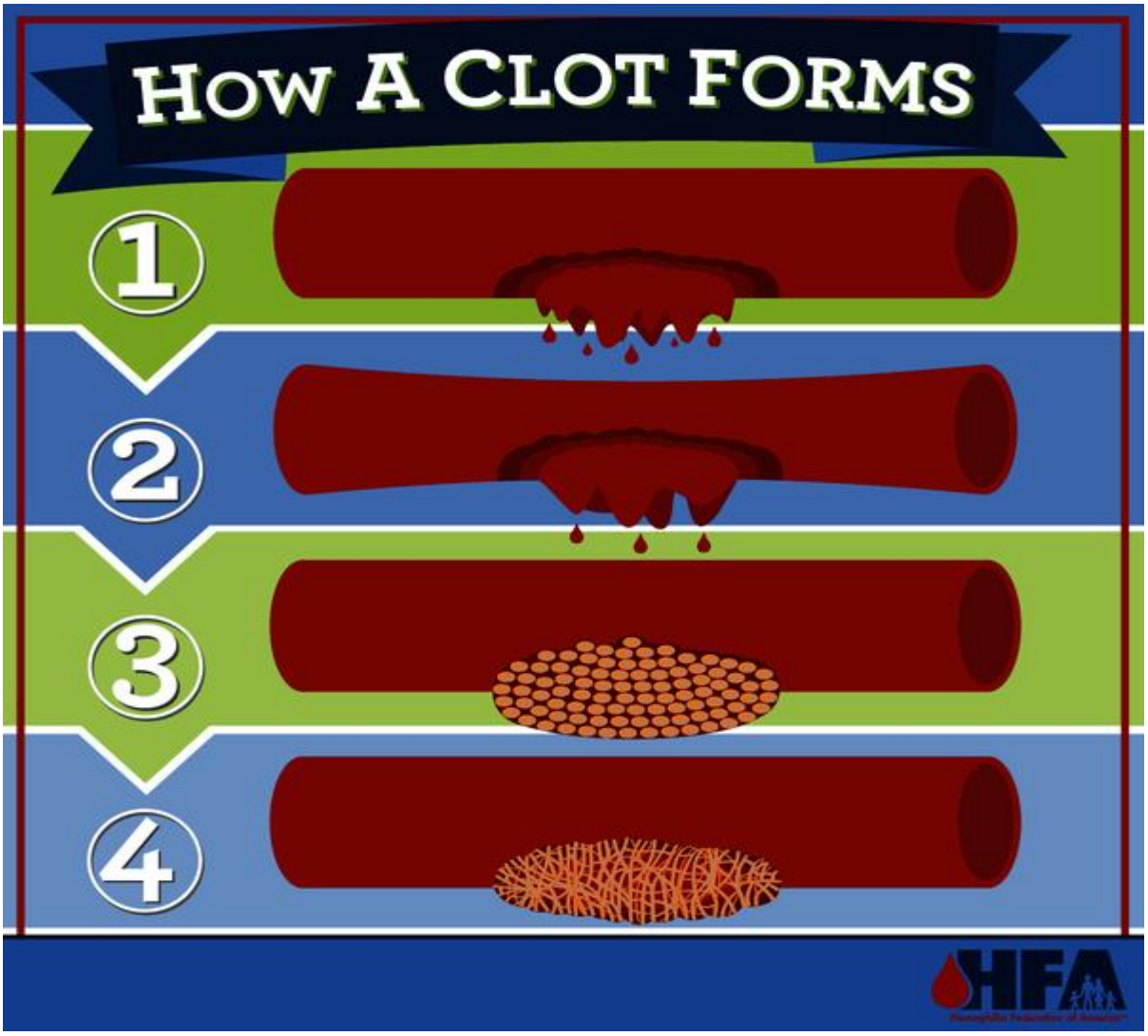
Haemophilia (= haemophilia) is caused by a sex-linked recessive gene h located in the X-chromosome. The gene h fails to produce necessary factor for quick clotting. A female becomes haemaphilic only when both the X-chromosomes carry the gene h (X^hX^h). Such females generally die before birth because of combination of these two recessive alleles which produce lethal condition. A female possessing only one allele for haemophilia (XX^h) appears to be normal as the normal gene in the X dominates the h. Such females are known as carriers. In case of male the recessive gene h on the X expresses itself as the Y chromosome is devoid of any corresponding allele (X^hY). The pedigree study of haemophilia was first made by Haldane in the royal families of Europe. The pedigree started from Queen Victoria in the last century. The ancestors of the queen did not have the disease. It appears that the gene for haemophilia developed either in the germ cells of her father or herself through mutation.



Molecular Causes:

Variants in the *F8* gene cause haemophilia A, while variants in the *F9* gene cause haemophilia B. The *F8* gene provides instructions for making a protein called coagulation factor VIII. A related protein, coagulation factor IX, is produced from the *F9* gene. Coagulation factors are proteins that work together in the blood clotting process. After an injury, blood clots protect the body by sealing off damaged blood vessels and preventing excessive blood loss.


Normal blood vessel
Platelets
Injured blood vessel
Blood clot
Activated platelets
Fibrin


How A Clot Forms
1
2
3
4
HFA

Variants in the *F8* or *F9* gene lead to the production of an abnormal version of coagulation factor VIII or coagulation factor IX, or reduce the amount of one of these proteins. The altered or missing protein cannot participate effectively in the blood clotting process. As a result, blood clots cannot form properly in response to injury. These problems with blood clotting lead to continuous bleeding that can be difficult to control. The variants that cause severe haemophilia almost completely eliminate the activity of coagulation factor VIII or coagulation factor IX. The variants involved in mild and moderate haemophilia reduce but do not eliminate the activity of one of these proteins.

Another form of the disorder, known as acquired haemophilia, is not caused by inherited gene variants. This rare condition is characterized by abnormal bleeding into the skin, muscles, or other soft tissues, usually beginning in adulthood. Acquired haemophilia results when the body makes specialized proteins called autoantibodies that attack and disable coagulation factor VIII. The production of autoantibodies is sometimes associated with pregnancy, immune system disorders, cancer, or allergic reactions to certain drugs. In about half of cases, the cause of acquired haemophilia is unknown

Like other sex-linked genes the gene for haemophilia shows criss-cross inheritance as follows:

A Carrier Woman marries a Normal man. A carrier Woman for haemophilia (XX^h) (Fig. 5.27) marries a normal man. The carrier woman produces ova of two types, one with X and other with X^h. The normal male also produces sperms of two types, one with X and other with Y chromosome.

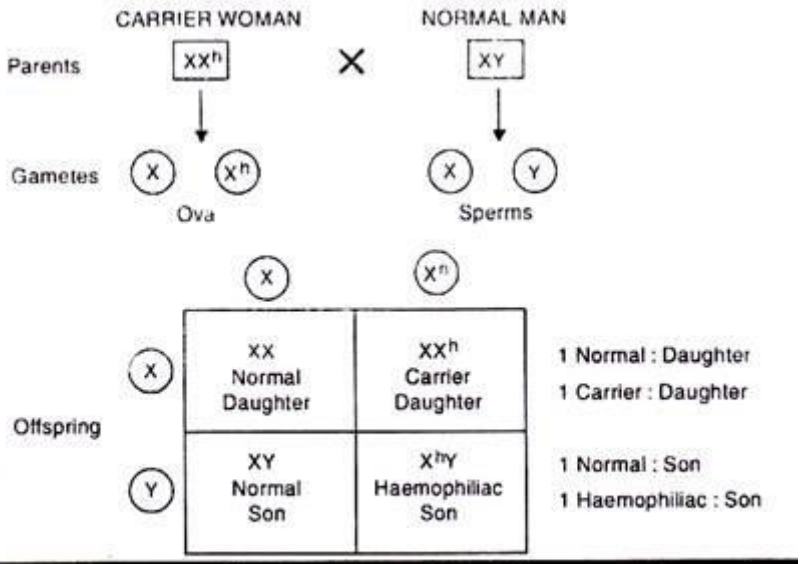


Fig. 5.27 Inheritance of haemophilia by 50% of the male children when the mother is carrier and the father is normal

The marriage can produce four types of children of combinations XX, XX^h, X^hY, XY (Fig. 5.27). Among the daughters 50% are normal and the remaining 50% are carriers.

Among the sons 50% are normal and rest 50% arc haemophilic. The carrier daughters are normal but transmits their haemophilic gene to 50% their children. If a haemophilic man (X^hY) marries a normal woman (XX) all the daughters are carriers as they receive one X^h from their lather whereas all the sons are normal as they receive X from their mother (Fig. 5.28).

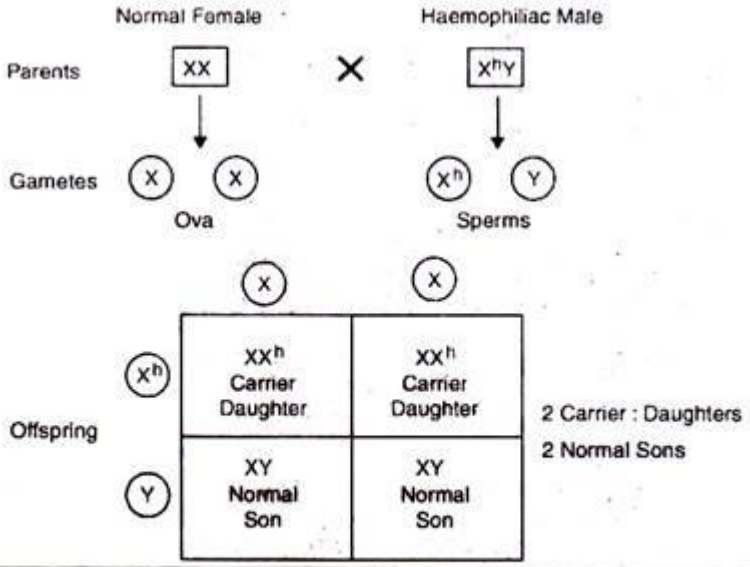


Fig. 5.28 Production of normal children when the father is haemophiliac and the mother is normal. This is because the X-chromosome of the male carrying the defective gene is passed on to the daughter who becomes carrier

A marriage between a carrier woman and a haemophilic man produces 50% sons normal and remaining 50% sons are haemophilic (Fig. 5.29). Among the daughters 50% are carriers and the remaining 50% are haemophilic and dies.

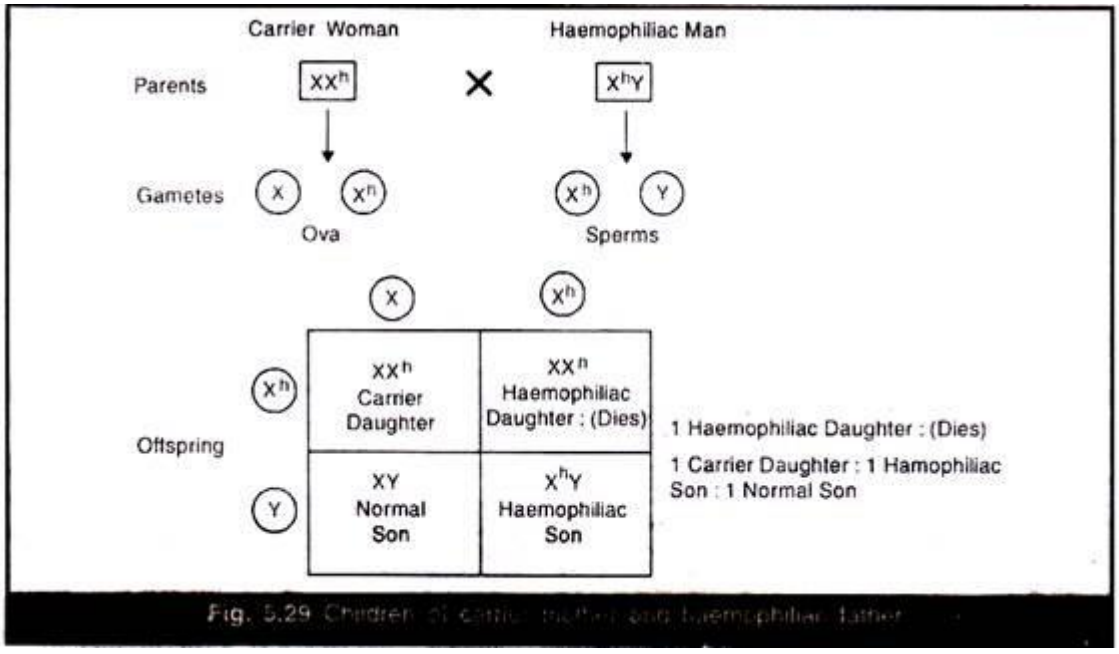


Fig. 5.29 Children of carrier mother and haemophiliac father

There are three types of haemophilia:

- **Haemophilia A:** This is the most common type of haemophilia. It happens when you don't have enough clotting factor 8 (factor VIII). The CDC estimates about 10 in 100,000 people have haemophilia A.
- **Haemophilia B:** Haemophilia B happens when you don't have enough clotting factor 9 (factor IX.) The CDC estimates about 3 in 100,000 people in the U.S. have haemophilia B.
- **Haemophilia C:** Haemophilia C is also known as factor 11 (factor XI) deficiency. This haemophilia type is very rare, affecting 1 in 100,000 people.

Haemophilia symptoms:

The most significant symptom is unusual or excessive bleeding or bruising.

People with haemophilia may develop large bruises after minor injuries. This is a sign of bleeding under their skin. They may bleed for an unusually long time, whether that's bleeding after surgery, bleeding after dental treatment or simply bleeding from a cut finger. They may start bleeding for no apparent reason, such as sudden bloody noses.

How much bruising or bleeding people have depends on whether they have severe, moderate or mild haemophilia:

People with severe haemophilia often have spontaneous bleeding or bleeding for no apparent reason. People with moderate haemophilia who have serious injuries may bleed for an unusually long time. People with mild haemophilia may have unusual bleeding, but only after major surgery or injury.

Other symptoms may include:

Joint pain from internal bleeding: Joints in your ankles, knees, hips and shoulders may ache, swell or feel hot to the touch.

Bleeding into your brain: People with severe haemophilia very rarely develop life-threatening bleeding into their brains. Brain bleeds may cause persistent headaches, double vision or make you feel very sleepy. If you have haemophilia and have these symptoms, get help right away.

What are haemophilia symptoms in babies and children?

Sometimes, babies assigned male at birth with haemophilia are diagnosed because they bleed more than usual after being circumcised. Other times, children develop symptoms a few months after they're born. Common symptoms include:

Bleeding: Babies and toddlers may bleed from their mouths after minor injuries, like bumping their mouths on a toy.

Swollen lumps on their heads: Babies and toddlers who bump their heads often develop goose eggs — large round lumps on their heads.

Fussiness, irritability or refusal to crawl or walk: These symptoms may happen if babies and toddlers have internal bleeding into a muscle or joint. They may have areas on their bodies that look bruised and swollen, feel warm to your touch or make your child hurt when you gently touch the area.

Hematomas: A hematoma is a mass of congealed blood that gathers under babies' or toddlers' skin. Babies and toddlers may develop hematomas after receiving an injection.

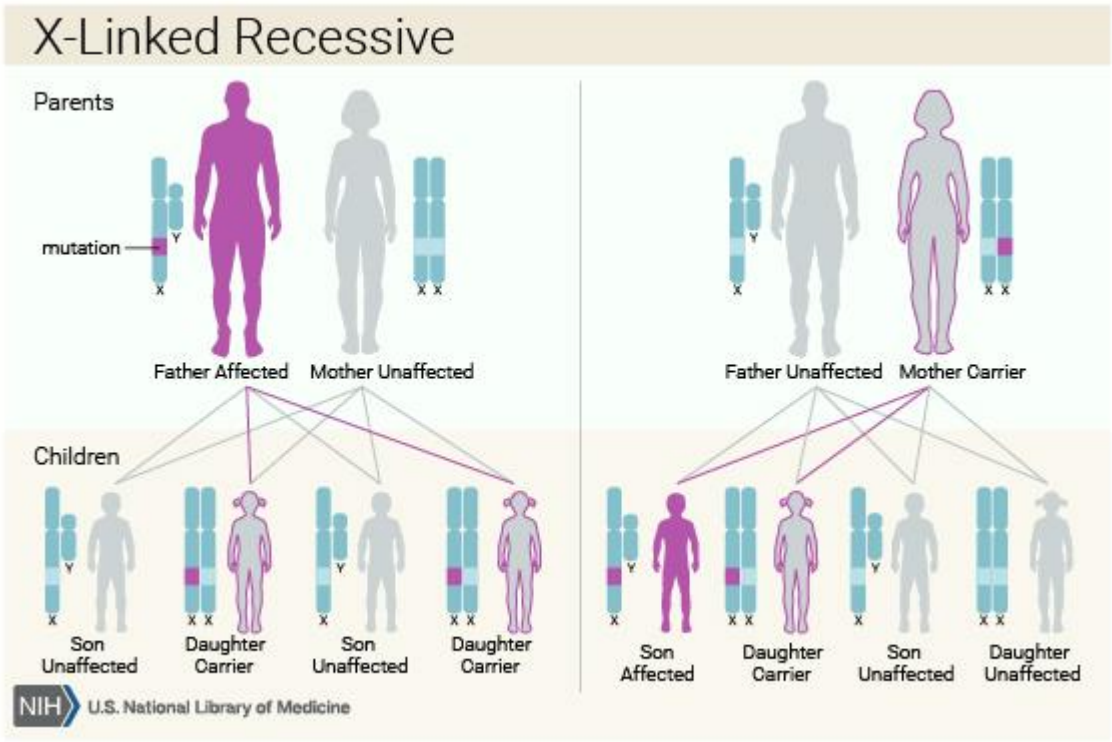
What causes haemophilia?

Certain genes create clotting factors. In inherited haemophilia, the genes carrying instructions for making normal clotting factors mutate or change. The mutated genes may give instructions that end up making abnormal clotting factors or not enough clotting factors. That said, about 20% of all haemophilia cases are spontaneous, meaning someone has the disease even though there's no family history of abnormal bleeding.

How do people inherit haemophilia?

Haemophilia A and B are both sex-linked disorders that are inherited in an X-linked recessive manner. Here's how that happens:

Everyone receives one set of chromosomes from their biological mother and one set of chromosomes from their biological father. If you get an X chromosome from your mother and an X chromosome from your father, you are assigned female at birth. If you get an X chromosome from your mother and a Y chromosome from your father, you are assigned male at birth. In other words, a mother will always pass an X chromosome to her offspring. The father will determine the assigned sex at birth by providing either an X or a Y chromosome. If a woman has an abnormal factor gene on one of their X chromosomes, they carry haemophilia but may not have symptoms. That's because there's a normal factor gene on their second X chromosome. If a woman who carries an X chromosome with a defective gene for producing factor 8 (or factor 9) has a male child, that child has a 50% chance of inheriting the X chromosome that carries the abnormal factor gene. If that same woman has a female child, that child has a 50% chance of inheriting the faulty chromosome and abnormal factor gene. That child likely wouldn't have symptoms because they'll also inherit a normal X chromosome from their father.



In other words, a woman who inherits a faulty X chromosome and abnormal factor gene will carry haemophilia. They may not have symptoms, but they can pass the condition on to their children. There's a 50% chance that any children they have — boys or girls — will inherit haemophilia. Boys who do inherit haemophilia are more likely to have severe symptoms. That's because they don't get a healthy X chromosome from their father. They can, but those symptoms tend to be mild. For example, a woman carrying the haemophilia gene may not have the normal clotting factors or not enough clotting factors. When that happens, they may have unusually heavy menstrual periods and they may bruise easily. They may bleed more after childbirth and they may develop joint problems if they have internal bleeding into their joints.

Hemophilia A and hemophilia B are inherited in an X-linked recessive pattern. The genes associated with these conditions are located on the X chromosome, which is one of the two sex chromosomes. In males (who have only one X chromosome), one altered copy of the gene in each cell is sufficient to cause the condition. A characteristic of X-linked inheritance is that fathers cannot pass X-linked traits to their sons.

In females (who have two X chromosomes), a variant would usually have to occur in both copies of the gene to cause the disorder. However, in some instances, one altered copy of the F8 or F9 gene is sufficient, because the X chromosome with the normal copy of the gene is turned off through a process called X-inactivation. X-inactivation occurs early in embryonic development in females. Through this process, one of the two X chromosomes is permanently turned off (inactivated) in somatic cells (cells other than egg and sperm cells). X-inactivation ensures that females, like males, have only one active copy of the X chromosome in each body cell.

Usually X-inactivation occurs randomly, such that each X chromosome is active in about half of the body cells. Sometimes X-inactivation is not random, and one X chromosome is active in more than half of cells. When X-inactivation does not occur randomly, it is called skewed X-inactivation.

In many females with a variant in one copy of the F8 or F9 gene, X-inactivation is random and the chromosome with the normal copy of the gene is turned off in about half of cells. These individuals have about half the usual amount of coagulation factor VIII or coagulation factor IX, which is generally enough for normal clotting. However, in some females with an F8 or F9 gene variant, X-inactivation is skewed, and the chromosome with the normal copy of the gene is turned off in more than half of cells. These individuals can have less coagulation factor VIII or coagulation factor IX than usual and are at risk of abnormal bleeding.

A healthcare provider will start by doing a complete history and physical examination. If you have haemophilia symptoms, the provider will ask about your family's medical history. Providers may do the following tests:

Complete blood count (CBC): Providers use this test to measure and study blood cells.

Prothrombin time (PT) test: Providers use this test to see how quickly your blood clots.

Activated partial thromboplastin time test: This is another blood test to time blood clot formation.

Specific clotting factor test(s): This blood test show levels of specific clotting factor levels (such as factor 8 and factor 9).

What are clotting factor levels?

Clotting factors help control bleeding. Healthcare providers categorize haemophilia as being mild, moderate or severe based on the amount of clotting factors in your blood:

People who have 5% to 30% of the normal amount of clotting factors in their blood have mild haemophilia. People with 1% to 5% of the normal level of clotting factors have moderate haemophilia. People with less than 1% of the normal clotting factors have severe haemophilia.

How do healthcare providers treat haemophilia?

They treat haemophilia by boosting clotting factor levels or replacing missing clotting factors (replacement therapy). In replacement therapy, you receive human plasma concentrates or lab-made (recombinant) clotting factors. In general, only people with severe haemophilia need regular replacement therapy. People with mild or moderate

haemophilia who need surgery may receive replacement therapy. They may also receive antifibrinolytics, a medication that keeps blood clots from breaking down.

Blood factor concentrates are made from donated human blood that's been treated and screened to reduce the risk of transmitting infectious diseases, such as hepatitis and HIV. People receive replacement factors via intravenous infusion (IV). If you have severe haemophilia and frequent bleeding episodes, your healthcare provider may prescribe prophylactic factor infusions to prevent bleeding.

The lifespan expectancy for someone with haemophilia:

According to 2012 data from the World Federation of Haemophilia, the lifespan for men and people AMAB with haemophilia is about 10 years fewer than the lifespan for men/people AMAB without haemophilia. The federation also states that children diagnosed with and treated for haemophilia have a normal life expectancy. But everyone is different. What's true for one person with haemophilia may not be true for others. If you or your child has haemophilia, ask your provider what you can expect. They know your/your child's situation, including overall health, and they're your best resource for information.

Activities and actions to improve quality of life

There are many things you can do to limit the impact haemophilia may have on your quality of life:

Develop an exercise routine: You may worry about hurting yourself during exercise. Talk to your provider about ways to reduce the risk of bleeding while staying active.

Manage your stress: Haemophilia is a lifelong illness. It may take extra effort to balance your obligations to your family and work.

Have good dental hygiene: Brushing, flossing and regular visits to your dentist reduce the risk that you'll need dental treatment that may cause bleeding. Be sure your dentist knows about your medical condition.

Aim for a weight that's healthy for you: Managing your weight may help if you're having trouble getting around because internal bleeding damaged your joints.

Educate those around you: If you have a severe form of haemophilia, you may have spontaneous bleeding that's hard to control even if you're taking medication. Make sure your family knows what to do if you have spontaneous bleeding. If your child has haemophilia, make sure care providers and school officials know what to do if your child has bleeding issues.

Phenylketonuria:

History:

PKU was first described by Asbjørn Følling, one of the first Norwegian physicians to apply chemical methods to the study of medicine. In 1934, the mother of two intellectually impaired children approached Følling to ascertain whether the strange musty odour of her children's urine might be related to their intellectual impairment. The urine samples were tested for a number of substances including ketones. When ketones are present, urine usually develops a red-brown colour upon the addition of ferric chloride, but in this instance the urine yielded a dark-green colour. After confirming that the unusual result was not due to any medications and repeating the test every other day for two months, Følling proceeded with a more detailed chemical analysis involving organic extraction and purification of the responsible compound, and determination of its melting point.[3] The basic elements were quantitated by combustion, and an empiric formula of $C_9H_8O_3$ derived.[3] Mild oxidation of the purified substance produced a compound which smelled of benzoic acid, leading Følling to postulate that the compound was phenylpyruvic acid. There was no change in the melting point upon mixing of the unknown compound with phenylpyruvic acid thus confirming the mystery compound was indeed phenylpyruvic acid.

Følling subsequently requested urine samples from 430 intellectually impaired patients from a number of local institutions and observed a similar result upon addition of ferric chloride, in a further eight individuals. These eight individuals all presented with a mild complexion (often with eczema), stooping figure with broad shoulders, a spastic gait, and severe intellectual impairment. Family studies of the affected individuals led to the suggestion of an inherited recessive autosomal trait. Dr Følling published his findings and suggested the name 'imbecillitas phenylpyruvica' relating the intellectual impairment to the excreted substance,[4] thereafter renamed 'phenylketonuria'.

Understanding has changed dramatically in the 70 years that have elapsed since the discovery of PKU. Jervis established the metabolic block and enzyme deficiency, and at about the same time, the link between reduced Phe intake and improved prognosis was shown.[8] After the birth of his intellectually impaired son and a niece with PKU, the Canadian Paediatrician Robert Guthrie, changed his research interests and developed screening tests for PKU. In the late 1970s, various groups began investigating the molecular basis of PKU. The most notable recent advance in the study of PKU was the establishment, in 1996, of the *PAH* Mutation Analysis Consortium Database.

The discovery of PKU by Dr Asbjørn Følling was an important milestone in medicine. The PKU model was used to illustrate how metabolic abnormalities could have neurological effects and how treatment could dramatically affect the clinical

manifestations of the disorder. The development of Guthrie's screening test, and dietary treatment, led to the prevention of intellectual impairment in affected children throughout the world. Futhermore, the PKU model has since been used as a template to shed light on over 200 other inborn errors of metabolism

PKU occurs due to a defect in the metabolism of amino acid phenylalanine. The affected children show normal physical development but impaired mental development to a varying degree. The untreated children become mentally defective adults. Phenylketonuria (PKU) is a rare genetic condition that causes an amino acid called phenylalanine to build up in the body. Amino acids are the building blocks of protein. Phenylalanine is found in all proteins and some artificial sweeteners.

Phenylalanine hydroxylase is an enzyme your body uses to convert phenylalanine into tyrosine, which your body needs to create neurotransmitters such as epinephrine, norepinephrine, and dopamine. PKU is caused by a defect in the gene that helps create phenylalanine hydroxylase. When this enzyme is missing, your body can't break down phenylalanine. This causes a buildup of phenylalanine in your body. Babies in the United States are screened for PKU shortly after birth. The condition is uncommon in this country, only affecting about 1 in 10,000 to 15,000 newborns each year. The severe signs and symptoms of PKU are rare in the United States, as early screening allows treatment to begin soon after birth. Early diagnosis and treatment can help relieve symptoms of PKU and prevent brain damage.

Biochemistry of PKU:

Phe exists as D and L enantiomers, and L-Phe is an essential amino acid required for protein synthesis in humans. Figure below shows the many processes which contribute to the flux of L-Phe in humans. As with many other metabolites, Phe concentrations are regulated to a steady state level with dynamic input and runout flux. Persistent disturbance to the flux will eventually result in alteration of the steady state concentrations. Dietary intake of Phe along with endogenous recycling of amino acid stores are the major sources of Phe, whereas, utilisation or runout of Phe occurs via integration into proteins, oxidation to Tyr, or conversion to other metabolites.

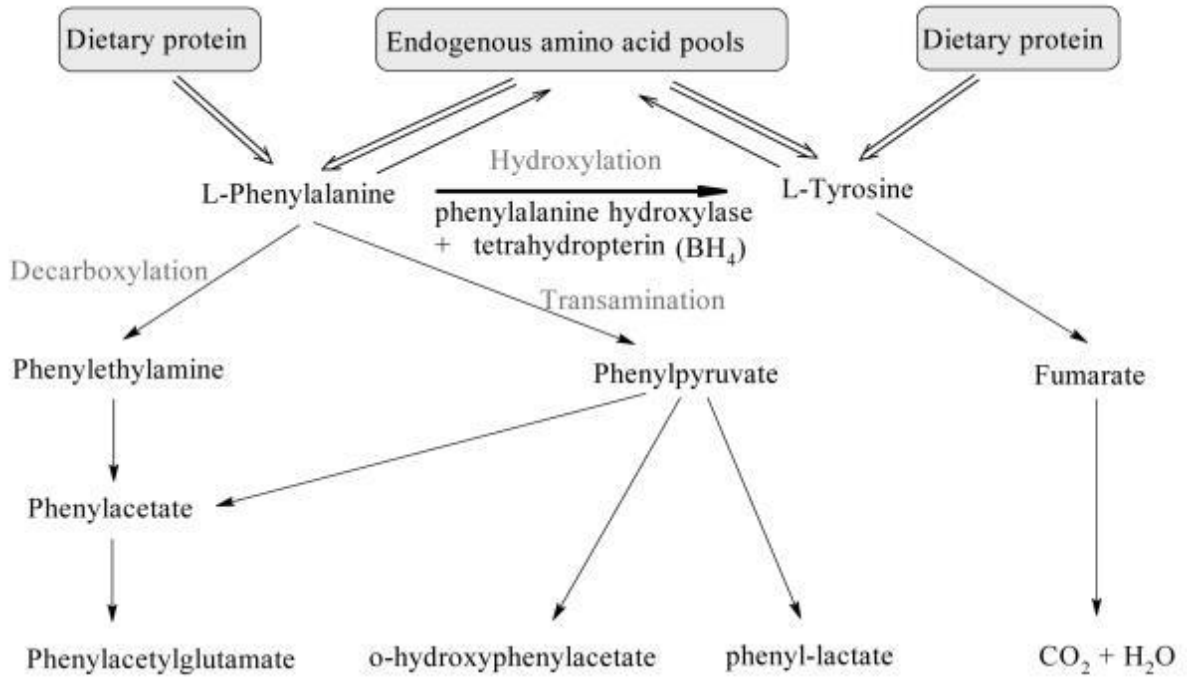


Figure: Phe metabolism in humans. Intake of l-Phe is via the diet and it is recycled through amino acid pools. Hydroxylation by PAH with its cofactor BH4, in the presence of molecular O2, produces L-Tyr. Alternative metabolism of L-Phe by decarboxylation or transamination produces various metabolites which are excreted in urine.

The conversion of Phe to Tyr occurs by a hydroxylating system consisting of:

a. **PAH**
b. **the unconjugated pterin cofactor, tetrahydrobiopterin (BH4)**
c. **enzymes which serve to regenerate BH4, namely dihydropteridine reductase and 4α-carbinolamine dehydratase**

While the para-hydroxylation of Phe is essential for the rupture of the benzene ring, it is not required for further metabolism of the alanine side chain.16 This alternative pathway of transamination and decarboxylation leads to the formation of metabolites such as phenylpyruvate, phenyllactate, and o-hydroxyphenylacetate which are excreted in urine. Conversion of Phe to Tyr (Figure 2) has two outcomes. First, it drives the endogenous production of the non-essential amino acid Tyr.16 Second, the hydroxylation reaction is the rate limiting step for complete oxidation of Phe to CO2 and H2O and contributes to the pool of glucose and 2-carbon metabolites.

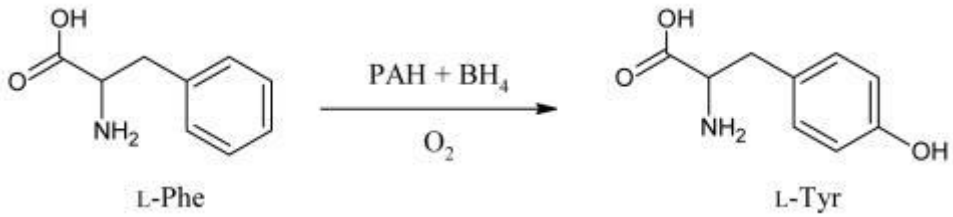


Figure: Conversion of Phe to Tyr is via a pathway involving the para-hydroxylation of the benzene by PAH, the cofactor BH4 and molecular oxygen.

A number of rare, related disorders due to defects in the BH_4 regeneration system can also affect Phe homeostasis, and catecholamine and serotonin biosynthesis, as this cofactor is common to the Phe, Tyr and tryptophan (Trp) hydroxylating enzymes.[18]

Phenylalanine Hydroxylase:

PAH catalyses the stereospecific hydroxylation of L-Phe, the committed step in the degradation of this amino acid. Phe catabolism and PAH activity is mainly associated with the liver, although minor activity has been demonstrated in rat kidney.[19] In humans, the PAH enzyme exists as a mixture of tetramers and dimers; the monomer is about 50 kDa in size and is comprised of 452 amino acids. The enzyme PAH requires BH_4 as a cofactor, as well as molecular oxygen for its activity.

PAH can be divided into a number of functional domains (Figure 3, Figure 4). The regulatory domain contains a serine residue which is thought to be involved in activation by phosphorylation. The catalytic domain contains a motif of 26 or 27 amino acids responsible for cofactor and ferric iron binding. The C-terminal domain is thought to be associated with inter-subunit binding.

PAH is regulated by a number of possible mechanisms. After a protein meal, it is postulated that the increased Phe in the amino acid pool causes a release of glucagon from the pancreas. Hepatic PAH is subject to control by cAMP-dependent protein kinase and α-andrenergic agent stimulated Ca^{2+}/calmodulin-dependent protein kinase phosphorylation dephosphorylation processes.[19] It has been further reported that these control mechanisms influence BH_4 co-factor interaction with PAH. In addition, there is evidence that Phe may also be able to cause a conformational change in PAH, as well as up-regulate cAMP activity. X-ray crystallographic studies are consistent with these mechanisms. Taken together, these mechanisms enable fine regulation of Phe concentrations by balancing levels sufficient for maintenance of protein biosynthesis while minimising tissue exposure to high concentrations of Phe.

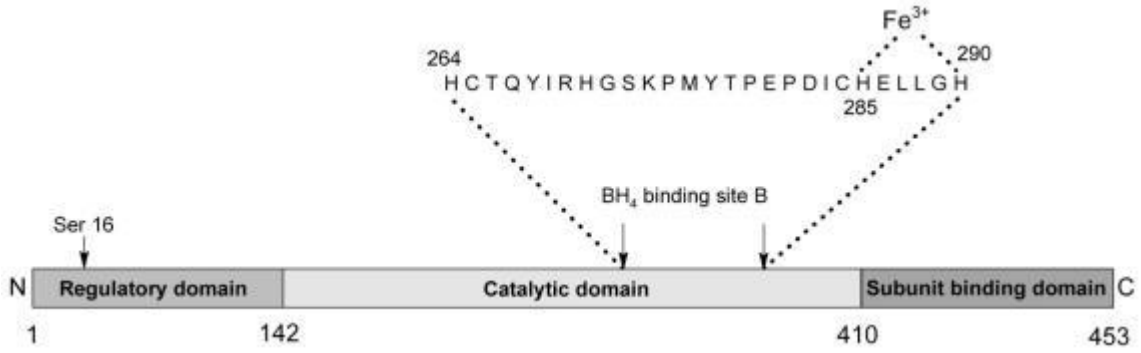


Figure: Structural components of PAH. The catalytic domain of PAH contains a motif of 26 or 27 amino acids which are responsible for ferric iron and cofactor (BH4) binding.

Symptoms of phenylketonuria

PKU symptoms can range from mild to severe. The most severe form of this disorder is known as classic PKU. An infant with classic PKU may appear normal for the first few months of their life. If the baby isn't treated for PKU during this time, they'll start to develop the following symptoms:

- seizures
- tremors, or trembling and shaking
- stunted growth
- hyperactivity
- skin conditions such as eczema
- a musty odor of their breath, skin, or urine

If PKU isn't diagnosed at birth and treatment isn't started quickly, the disorder can cause:

- irreversible brain damage and intellectual disabilities within the first few months of life
- behavioral problems and seizures in older children

A less severe form of PKU is called variant PKU or non-PKU hyperphenylalaninemia. This occurs when the baby has too much phenylalanine in their body. Infants with this form of the disorder may have only mild symptoms, but they'll need to follow a special diet to prevent intellectual disabilities.

Once a specific diet and other necessary treatments are started, symptoms start to diminish. People with PKU who properly manage their diet usually don't show any symptoms.

Causes of phenylketonuria

PKU is an inherited condition caused by a defect in the PAH gene. The PAH gene helps create phenylalanine hydroxylase, the enzyme responsible for breaking down phenylalanine. A dangerous buildup of phenylalanine can occur when someone eats high-protein foods, such as eggs and meat. Both parents must pass on a defective version of the PAH gene for their child to inherit the disorder. If just one parent passes on an altered gene, the child won't have any symptoms, but they'll be a carrier of the gene.

Diagnosis of PKU:

Since the 1960s, hospitals in the United States have routinely screened newborns for PKU by taking a blood sample. A doctor uses a needle or lancet to take a few drops of blood from your baby's heel to test for PKU and other genetic disorders. The screening test is performed when the baby is one to two days old and still in the hospital. If you don't deliver your baby in a hospital, you'll need to schedule the screening test with your doctor. Additional tests may be performed to confirm the initial results. These tests search for the presence of the PAH gene mutation that causes PKU. These tests are often done within six weeks after birth. If a child or adult shows symptoms of PKU, such as developmental delays, the doctor will order a blood test to confirm the diagnosis. This test involves taking a sample of blood and analyzing it for the presence of the enzyme needed to break down phenylalanine.

Treatment options:

People with PKU can relieve their symptoms and prevent complications by following a special diet and by taking medications.

Diet

The main way to treat PKU is to eat a special diet that limits foods containing phenylalanine. Infants with PKU may be fed breast milk. They usually also need to consume a special formula known as Lofenalac. When your baby is old enough to eat

solid foods, you need to avoid letting them eat foods high in protein. These foods include:

- eggs
- cheese
- nuts
- milk
- beans
- chicken
- beef
- pork
- fish

To make sure that they still receive an adequate amount of protein, children with PKU need to consume PKU formula. It contains all the amino acids that the body needs, except for phenylalanine. There are also certain low-protein, PKU-friendly foods that can be found at specialty health stores. People with PKU will have to follow these dietary restrictions and consume PKU formula throughout their lives to manage their symptoms.

It's important to note that PKU meal plans vary person to person. People with PKU need to work closely with a doctor or dietitian to maintain a proper balance of nutrients while limiting their intake of phenylalanine. They also have to monitor their phenylalanine levels by keeping records of the amount of phenylalanine in the foods they eat throughout the day. Some state legislatures have enacted bills that provide some insurance coverage for the foods and formulas necessary to treat PKU. Check with your state legislature and medical insurance company to find out if this coverage is available for you. If you don't have medical insurance, you can check with your local health departments to see what options are available to help you afford PKU formula.

Medication:

The United States Food and Drug Administration (FDA) recently approved sapropterin (Kuvan) for the treatment of PKU. Sapropterin helps lower phenylalanine levels. This medication must be used in combination with a special PKU meal plan. However, it doesn't work for everyone with PKU. It's most effective in children with mild cases of PKU.

Pregnancy and phenylketonuria:

Woman with PKU may be at risk of complications, including miscarriage, if they don't follow a PKU meal plan during their childbearing years. There's also a chance that the unborn baby will be exposed to high levels of phenylalanine. This can lead to various problems in the baby, including:

- intellectual disabilities
- heart defects
- delayed growth
- low birth weight
- an abnormally small head

These signs aren't immediately noticeable in a newborn, but a doctor will perform tests to check for signs of any medical concerns your child may have.

i. The Biochemical and Genetic Defects:

a. PKU had been named by the finding of a ketone, phenyl pyruvic acid in the urine. The primary defect is in the phenylalanine hydroxylating system which converts the amino acid into tyrosine.

b. This system when cannot convert all of the phenylalanine derived from the protein in a mother's milk, the level in the blood rises and this causes the impaired development of the nervous system. Some of the excess phenylalanine is deaminated to phenyl pyruvic acid which is excreted in the urine.

c. Hydroxylation is effected by phenylalanine hydroxylase (PH) and a coenzyme 5, 6, 7, 8-tetra-hydrobiopterin (BH_4). The coenzyme is oxidized to BH_2 from which it is reformed by another enzyme, di-hydrobiopterin reductase (DHPR). Hydroxylation may be impaired by genetic defects in production of pH of DHPR and of an enzyme responsible for the formation of BH_2.
d. The genes responsible for the control of the hydroxylation are not closely linked and may be on different chromosomes. The defects are transmitted by autosomal recessive inheritance. Unaffected heterozygous individuals act as carriers.

ii. Classical Phenylketonuria:

a. When the baby is at the age of 8 to 10 months, the parents may become anxious because their child is slow in learning to sit and handle things and is generally un-responsive.

b. About 25 p.c. of the affected children develop eczema.

c. The retarded development becomes obvious and there may be signs of severe birth damage, such as myoclonic epilepsy and marked hyperactivity.

d. Most affected children grow up to become physically sound but are mentally defective.

iii. Variant Forms:

Besides classical PKU, other forms of hyper-phenyl-alanemia are known and at least nine types have been recorded.

iv. Dietary Management:

a. Clinical symptoms do not arise in case the affected infant is put on a low phenylalanine diet soon after birth and kept on it for a long period.

b. A severe emotional strain is imposed on a young child on an entirely artificial diet.

c. As soon as the diagnosis is made, breast feeding should be stopped and the infant is bottle-fed with a low phenylalanine milk substitute.

d. Greater difficulty arises when the baby has to be weaned. A mother then has to prepare a low phenylalanine diet for her child from five lists of foods. Therefore, she needs continuing help from a dietician.

The lists are:

(i) Basic foods containing negligible phenylalanine which can be used freely (these include sugar, sweets, jams, solid vegetable oils and cooking oils).

(ii) Fruits and vegetables which can be taken freely since they provide negligible phenylalanine and protein in a normal helping.

(iii) A basic list of 59 mg phenylalanine exchanges of foods.

(iv) Manufacturer's foods of negligible phenylalanine content.

(v) Exchanges of foods containing 50 mg of phenylalanine (by calculation taking one gram of protein as 50 mg phenylalanine).

Emerging treatment strategies of PKU:

Although dietary restriction of Phe is the cornerstone of treatment for PKU, the practicalities of following the strict diet have led to trials of additional therapies.

BH_4 Therapy:

Recent clinical trials have shown that a subset of 'classical' PKU children respond to BH_4 therapy, dependent upon their PAH gene mutation(s).[70] Sapropterin dihydrochloride (Kuvan, Biomarin Pharma) is an orally active synthetic form of BH_4 that has received Orphan Drug status and Fast Track designation for the treatment of PKU. Phase II and III clinical trials have shown that Kuvan is a safe and effective therapy in selected patients with HPA and mild-to-moderate PKU who responded to a BH_4 loading test.

Enzyme Replacement Therapy:

Unfortunately, patients with more severe forms of classic PKU and some non-PKU HPA do not respond to BH_4 treatment, presumably because these individuals lack sufficient residual PAH activity for stimulation by BH_4. Such nonresponders could benefit most from enzyme replacement therapy. Unlike BH_4 treatment, enzyme replacement is not dependent on the *PAH* genotype. Replacement of the enzyme could be facilitated by partial liver or normal hepatocyte transplantation. Although liver transplantation would correct the metabolic phenotype in PKU, the high risk of major surgery and lifelong immunosuppression precludes its routine use.

An alternative enzyme therapy for PKU has been trialled which involves the substitution of PAH with Phe ammonia-lyase (PAL, EC 4.3.1.5), a non-cofactor dependent plant protein involved in Phe degradation. This treatment has been shown to be effective in mouse models causing modest but short-lived falls in Phe concentrations. Indeed, Phenylase™ (PAL), Biomarin Pharma is currently under investigation for the potential treatment of patients with PKU who do not respond to BH_4.

Large Neutral Amino Acid Therapy:

As discussed previously, it is hypothesised that competition for the L-type amino acid carrier by Phe with other LNAAs may occur in PKU. This hypothesis has led to LNAA supplementation trials. Increasing the blood concentrations of various LNAAs has led to reduced brain concentrations of Phe. Furthermore, the increased Tyr and Trp intake may be of benefit in disorders of BH_4 regeneration. A new LNAA formulation (NeoPhe, Solace Nutrition) has been effective in reducing blood Phe concentrations.

Gene Therapy:

Complete and persistent correction of HPA has been reported in the *Pah*^{enu2} mouse using somatic gene therapy, site specific genome integration of *Pah* cDNA and by liver directed recombinant adeno-associated virus vectors. Early work on gene therapy for

children with PKU was considered inappropriate as the therapy involved administration of immunosuppressant agents to block the immune response to the vector so as to prolong the therapeutic effect.[76] These trials involved the use of recombinant adenoviral vectors. Other trials involving the use of recombinant retroviral vectors have been abandoned following the observation that these vectors may induce leukaemia-like disorders.

PAH-deficient (*Pah*^{enu2}) mouse models have been used to trial gene transfer via a recombinant adeno-associated virus vector encoding the human *PAH* gene with promising results. This vector appears to be a safer mode of transfer as it possesses minimal antigenicity and showed no signs of inducing liver damage. However, administration of the vector would ideally be facilitated by a less intrusive means than via the hepatic portal vein, which was used in the aforementioned study.

Maternal PKU: Highly elevated concentrations of Phe are teratogenic and are a cause of increased risk of miscarriage. Specifically, the foetus can be affected by elevated Phe concentrations which lead to intrauterine growth retardation, facial dysmorphism, microcephaly, congenital heart disease and developmental delay and has led to families consisting of multiple children with birth defects and intellectual impairment. Foetal exposure to HPA in PKU affected pregnancies is exacerbated by the transplacental gradient for Phe; an average foetal/ maternal ratio of 1.5 is suggested although ratios up to 2.9 can be seen in early pregnancy when foetal development is maximal.[15] Because of the foetal morbidities associated with HPA, a more stringent control of blood Phe concentrations is required in mothers with PKU. A preconception diet is required with a Phe target interval of between 100 and 360 μmol/L in affected mothers. In addition, weekly monitoring of the Phe concentrations is advised to aid in achieving low baseline levels.

Conclusion: PKU is a historically significant inborn error of metabolism. Its discovery over 70 years ago and scientific investigation has established the link between metabolic disease and intellectual impairment, led to the development of neonatal screening programs across the globe, and demonstrated how effective treatment can lead to a near normal outcome for affected individuals. However, despite the intensive study, the mechanism by which the aberrant Phe metabolism leads to intellectual impairment is yet to be explained. The successes of the PKU story lie squarely in the hands of the health professionals who diagnose PKU through establishing and managing neonatal screening programs, and the paediatricians, dieticians and carers who then supervise PKU treatment. The study of PKU has revealed genomic components of health and disease. A better understanding of the biochemistry, genetics and molecular basis of PKU, as well as the need for improved treatment options, has led to the development of new therapeutic strategies.

Probable Questions:

1. Write down the characteristics of cystic fibrosis.
2. Describe inheritance pattern of Cystic fibrosis.
3. What are the symptoms of Cystic fibrosis?
4. How Cystic fibrosis can be diagnosed?
5. State the treatment strategies for cystic fibrosis.
6. Why haemophilia is caused?
7. Describe inheritance pattern of haemophilia.
8. Describe variation in haemophilia.
9. What are the symptoms of haemophilia?
10. State the treatment strategies for haemophilia.
11. Describe the biochemical basis related to Phenylketonuria.
12. What is the role of Phenylalanine hydroxylase in Phenylketonuria?
13. What are the symptoms of Phenylketonuria?
14. How PKU can be diagnosed?
15. State the treatment strategies for PKU.
16. What is classical PKU?
17. Describe some modern treatment strategies f PKU?
18. What is maternal PKU?
19. How PKU is related to pregnancy?
20. What are the variant forms of PKU?

Suggested readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics . Verma and Agarwal.

UNIT-VI

Human Genome: Human genome project and the age of genomics, Structure of Human Genome. Concepts and application of Bioinformatics

Objective: In this unit you will learn about different aspects of human genome project and about genomics. We will also discuss Concepts and application of Bioinformatics.

Meaning of Genomics:

The term genomics was first used by Thomas Roderick in 1986. It refers to the study of structure and function of entire genome of a living organism. Genome refers to the basic set of chromosomes. In a genome, each type of chromosome is represented only once. Now genomics is being developed as a sub discipline of genetics which is devoted to the mapping, sequencing and functional analysis of genomes.

Main points related to genomics are listed below:

i. It is a computer aided study of structure and function of entire genome of an organism.

ii. It deals with mapping and sequencing of genes on the chromosomes.

iii. It is a rapid and accurate method of gene mapping. It is more accurate than recombination mapping and deletion mapping techniques.

iv. The genomic techniques are highly powerful, efficient and effective in solving complex genetic problems.

v. Now use of genomic techniques has become indispensable in plant breeding and genetics.

Types of Genomics:

The discipline of genomics consists of two parts, viz. structural genomics and functional genomics.

These are defined as under:

i. Structural Genomics:

It deals with the study of the structure of entire genome of a living organism. In other words, it deals with the study of the genetic structure of each chromosome of the genome. It determines the size of the genome of a species in mega-bases [Mb] and also the genes present in the entire genome of a species.

ii. Functional Genomics:

The study of function of all genes present in the entire genome is known as functional genomics. It deals with transcriptome and proteome. The transcriptome refers to complete set of RNAs transcribed from a genome and proteome refers to complete set of proteins encoded by a genome.

3. Classification of Genomics:

The genomics can be classified as plant genomics, animal genomics, eukaryotic genomics and prokaryotic genomics.

These are defined as follows:

(i) Plant Genomics:

It deals with the study of structure and function of entire genome of plant species.

(ii) Animal Genomics:

It deals with the study of structure and function of entire genome of animal species.

(iii) Eukaryotic Genomics:

It deals with the study of structure and function of entire genome of higher [multi-cellular] organisms.

(iv) Prokaryotic Genomics:

It deals with the study of structure and function of entire genome of unicellular organisms.

Whole Genome Sequence Data:

Complete nucleotide sequences of nuclear, mitochondrial and chloroplast genomes have already been worked out in large number of prokaryotes and several eukaryotes. By the year 2005, among prokaryotes, approx. 1400 viral genomes, 250 bacterial genomes (230 eubacteria and 20 archaea), 500 mitochondrial genomes, 35 chloroplast genomes have been fully sequenced.

Among the eukaryotes namely the whole genome of *Saccharomyces cerevisiae* (yeast), *Coenorhabditis elegans* (nematode), fruitfly (*Drosophila melanogaster*), Human (*Homo sapiens*), Crucifer weed (Arabidopsis thaliana) and rice (*Oryza sativa*) have been sequenced already and data available for annotation studies.

The sequence data of eukaryotic nuclear genome is an important source of identi-fication, discovery and isolation of important genes. This data is very much helpful in variety of application relevant to animal, plant and microbial biotechnology.

Functional Genomics:

Functional genomics is to place all of the genes in the genome of an organism within a functional frame work. Actually, in every organism about 12-15% genes are structural genes which are expressed for certain characters. These are transcribed in a given cell. This is helpful in overall functioning of the cell and organism.

Functional genomics brings together genetics with gene transcripts, proteins and metabolites by analyzing genome sequencing. Functional genomics is driving a shift from vertical analysis of single genes, proteins or metabolites towards horizontal analysis of full suites of genes, proteins and metabolites. This may help in molecular participation of a given biological process. This offers the prospect of determining a truly holistic picture of life.

Functional Genomics Toolbox:

The functional genomics emerged in response to the challenges posed by complete genome sequences. To understand this process it is necessary to know the biochemical and physiological function of every gene product and their complexes.

The activity of genes manifests at a number of different levels, including RNA, protein and metabolite levels and analyses at these levels can provide insight not only into the possible function of individual gene but also the cooperation that occurs between genes and gene products to produce a defined biological outcome.

The technology, involved in defining functional genomics are DNA or oligonucleotide microarray technology for determining mRNA, 2D gels and mass spectroscopy and other methods for analyzing different proteins and GC-MS or LC-MS for identifying and quantifying different metabolites in a cell. High throughput methods for forward and reverse genetics are also integral to functional genomics (Fig. 27.25).

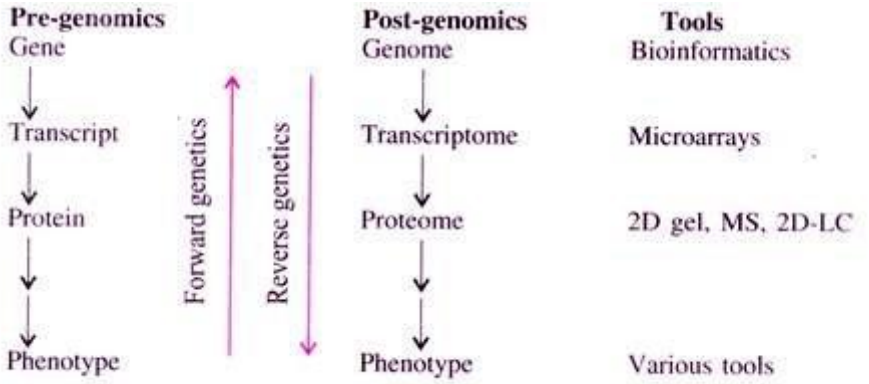


Fig. 27.25 : Tools of functional genomics.

Methods in Functional Genomics:

Functional genomics lies on gene expression, profiling (mRNA) in protein expression, reverse genetics, the generation of targeted mutations in genes of interest besides forward mutation rate, the generation of random mutations in the genome for desirable mutants and bioinformatics.

These criteria help in providing maximum information of a particular organism. This helps in understanding the biological process at the molecular level and also useful to identify novel genes regulating this process. To understand the gene function, it is desirable to identify genes and to understand its expression at the whole genome level.

There are many prokaryotic and eukaryotic organisms whose genomes are fully sequenced. The current discovery is mapping of whole sequences of genes present in human genome. It is possible to assign functions to novel genes and proteins, and to understand biological processes at the molecular level.The integrated understanding of the control of gene expression and knowledge of signal transduction, cell signalling and overall cell function are dynamic tools to study regulation of gene expression in any given cell type. In yeast cells, transcripts associated with different parts of the cell cycle form discrete clusters.

These studies allowed sequence tags encoding proteins of unknown function to be assigned to putative classes based on their clustering with genes of known function. Here, role of functional genomics will be to test those repetitive functions and apply to resolve complex biological processes.

Future Prospects:

It has good future as briefly given below:

(a) In Human Pathology:

Application of gene expression profiling in understanding the human cells and tissues to disease is under way. It would be possible in future to study the modifications in gene sequence during infection. Such studies will yield fundamental insights into the etiology and pathogenesis.

(b) In Parallel Sequencing:

In a number of laboratories, most of the sequences are generated using different approaches. Integrating very different datasets is not as simple as assembling a sequence itself which serves as an absolute standard. Although transcriptional profiling can be used to construct the standardized databases based on absolute RNA and protein levels, yet this is clearly not the case for relative gene expression data.

Significance of Genomics:

All the information's require input in probability theory, database management arid manipulation, and computer science.

This will help in:

(a) Identification of open reading frame sequences,

(b) Gene splicing sites (introns),

(c) Gene annotation (inter-genomic comparisons) and

(d) Determination of sequence patterns of regulatory sites and gene regulations.

Human Genome Project:

Introduction:

The term genome (introduced by H. Winkter in 1920) refers to one complete copy of the genetic information (DNA) or one complete set of chromosome (monoploid or haploid) of an organism. The term genomics (term coined by T.H. Roderick in 1987) denotes mapping, sequencing and functional analysis of genomes.

The Human Genome Project (HGP) was launched on 1st October, 1990 for sequencing entire genome of 2.75 billion nucleotide pairs. This megaproject was a 13 year project coordinated by the U.S. Department of Energy and the National Institute of Health. During the early years of the HGP, the Wellcom Trust (U.K.) becomes a major partner; additional contributions come from Japan, France, Germany, China and others.

The project was completed in 2003. James Watson was the first directors of human genome project and after a period of two years he was replaced by Francis Collins in 1993. Two important scientists associated with HGP were Francis Collins, director of the HGP and J. Craig Venter, founding president of Celera Genomics. HGP was closely associated with rapid development of a new era in biology called as Bioinformatics.

Meaning of Human Genome Project:

The Human Genome Project (HGP) is an International collaborative research programme which started in 1990 and completed in 2003, whose goal was the complete

mapping and understanding of the three billion DNA subunits (bases), and to identify all human genes, making them accessible for further biological study.

History of Human Genome Project:

In U.S., the HGP was carried out by the Department of Energy (Human Genome Program) directed by Ari Patrinos, and National Institute of Health (Human Genome Research Institute) directed by Francis Collins. In 2001, Craig Venter, CEO of Celera Genomics, co-announced the completion (90%) of sequencing of the human genome (draft sequence).

The full sequence was completed and published in 2003 (finished sequence). More refined sequence is available in 2006 and correction of minor errors (1 less in 10000 DNA subunits) requires some time to come.

The Birth and Activity of Human Genome Project:

The human genome project (HGP) was conceived in 1984, and officially begun in earnest in October 1990. The primary objective of HGP was to determine the nucleotide sequence of the entire human nuclear genome. In addition, HGP was also entrusted to elucidate the genomes of several other model organisms e.g. Escherichia coli, *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (roundworm), *Mus musculus* (mouse). James Watson (who elucidated DNA structure) was the first Director of HGP.

In 1997, United States established the National Human Genome Research Institute (NHGRI). The HGP was an international venture involving research groups from six countries—USA, UK, France, Germany, Japan and China, and several individual laboratories and a large number of scientists and technicians from various disciplines. This collaborative venture was named as International Human Genome Sequencing Consortium (IHGSQ and was headed by Francis Collins.

A total expenditure of $3 billion, and a time period of 10-15 years for the completion of HGP was expected. A second human genome project was set up by a private company — Celera Genomics, of Maryland USA in 1998. This team was led by Craig Venter. Very rapid and unexpected progress occurred in HGP with good cooperation between the two teams of workers and improved methods in sequencing.

Goals of Human Genome Project:

Human Genome Project had many goals some of the important goals were outlined below:

1. To identify all the approximately 20,000-25,000 genes in human DNA.

2. To determine the sequences of the 3 billion base pairs that makes up human DNA.

3. To store this information in data base.

4. To develop improvised tools for data analysis.

5. To transfer related technologies to other sectors, such as industries.

6. To address the ethical, legal and social issues (ELSI) that may arise from the project.

The methodologies involved two major approaches identifying all genes of the genome and their sequencing. For sequencing, the total DNA from a cell is isolated and converted into fragments of relatively small sizes and cloned in suitable host, this generates a genomic library of the organism. The complete sequencing of the first human chromosome, small chromosome 22, was published in December 1999. Then chromosome 21 was completely sequenced in May 2000. The first draft sequence of entire human genome was published in the famous scientific journal "Nature" on 16th February, 2001.

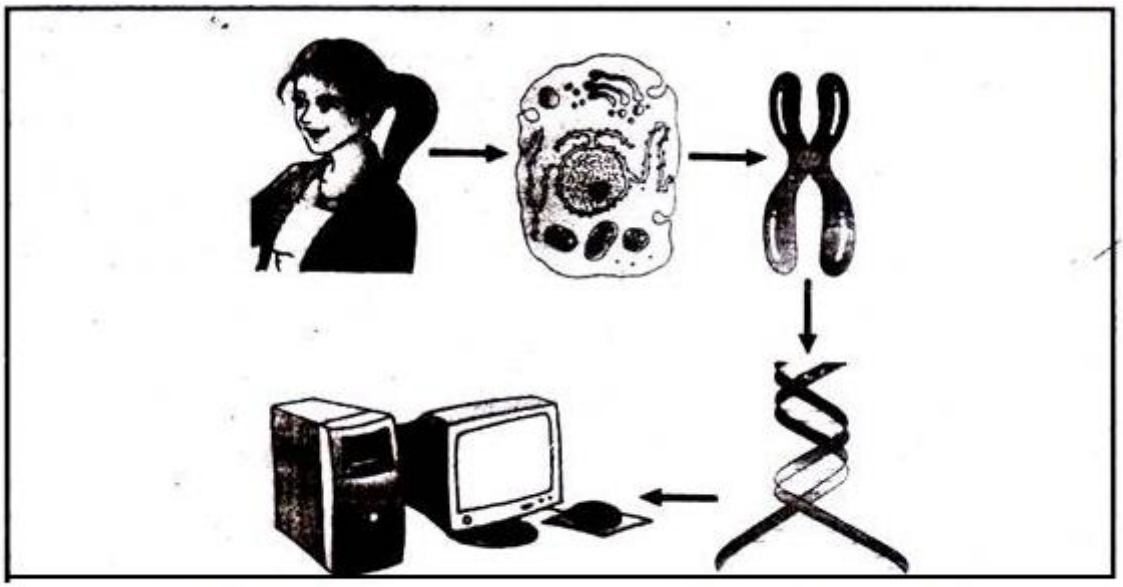


Fig 5.23 A Representative Diagram of Human Genome Project

Important Features of Human Genome:

1. The human genome contains 3164.7 million nucleotide bases.

2. The average gene consists of 3000 bases, but gene size vary greatly (the largest human gene is dystrophin containing 2.4 million bases).

3. The total number of genes in the genome is estimated at 30,000 and all (99.9 percent) nucleotide bases are exactly the same in all people

4. Functions of about 50% of the discovered genes are still unknown.

5. Less than 2% of the genes of the genome codes for proteins.

6. Chromosome 1 has most genes (2968) and the Y has the fewest (231).

7. Repeated sequences (AT-AT-AT or GC-GC-GC…….) make up very large portion of the human genome.

8. Scientists have identified about 1.4 million locations where single base DNA differences (SNPs-single nucleotide polymorphism, pronounced as 'snips') occur on humans.

Human Genome Project was an undertaking by many countries to acquire complete knowledge of the organisation, structure and functions of the human genome. Such a multinational undertaking was called as International Human Genome Sequencing. HGP was regarded as the most ambitious project ever undertaken by humans.

The project had the benefits to identify all human genes and also mutated genes causing diseases. Complete knowledge on the genome sequence will enable the scientists in future to gain knowledge on the types of proteins encoded by these genes. The cloning and sequencing of the disease causing alleles (mutated genes) will largely help in the diagnosis and treatment of diseases.

Human Genome Size:

A genome is an organism's complete set of deoxyribonucleic acid (DNA), a chemical compound that contains the genetic instructions needed to develop and direct the activities of an organism. The human genome contains approx. Three billion base pairs which reside in 23 pairs of chromosomes.

Each chromosome contains hundreds and thousands of genes, and ranges in size from about 50000000 to 300000000 base pairs. The total number of genes is 30000 (approx.) and accounts for only 25% of the DNA; the rest is extra-genic DNA.

Human Genome Project Mapping:

Before beginning a sequencing project of the human genome, it was first necessary to produce a good framework map. Two general methods were developed for mapping human genome — standard method and whole genome short-gun method.

The standard method involves finding a segment of the genome and locating where it belongs. Genetic maps based on recombination frequencies between markers are useful in ordering genes. Molecular markers like RFLP, VNTRs (Microsatellites), STSs, SNPs have been used in mapping human genome. The whole genome shotgun sequencing method involves shearing of genomic DNA followed by cloning, to produce a genomic library.

This is followed by sequencing of cloned DNA fragments at random, followed by shotgun assembly, i.e., the assembly of the fragment sequences into larger units on the basis of their overlaps. Groups of cloned DNA segments that can be aligned in an overlapping fashion to cover a region of the human genome are referred as contigs. Yeast Artificial Chromosomes (YACs) were initially used as cloning agents when primary task was mapping. However, as the emphasis of the project shifted to sequencing. Bacterial Artificial Chromosomes (BACs) were used.

Human Genome Project Sequence:

Sequencing means determining the exact order of the base pairs in a segment of DNA. The primary method used by the HGP to produce the finished version of the human genetic code is map-based or BAC- based sequencing. The human DNA is fragmented into pieces that are relatively large, cloned in the bacteria, stored for replication as required.

A collection of BAG clones containing the entire human genome is called a BAC-library. In this method, each BAC clone is mapped to determine the location of that fragment in human chromosome and then the DNA letters are sequenced from each clone and their spatial relation to sequenced human DNA in other BAC clones.

For sequencing, each BAC clone is cut into still smaller fragments that are about 2000 bases in length. These pieces are called **"sub-clones".** A **"sequencing reaction"** is carried out on these sub-clones. With the help of a computer then the short sequences are assembled into contiguous stretches of sequence of the clones.

In a short the whole process can be summarized:

i. Chromosomes, which range in size from 50 million to 250 million bases, must first be broken into much shorter pieces (sub-cloning step).

ii. Each short piece is used as a template to generate a set of fragments that differ in length from each other by a single base that will be identified in a later step (template preparation and sequencing step).

iii. The fragments in a set are separated by gel electrophoresis (separation step).

iv. The final base at the end of each fragment is identified (base-calling step). This process recreates the original sequence of As, Ts, Cs and Gs for each short piece generated in the first step.

v. After the bases are 'read', computers are used to assemble the short sequences (in blocks of about 500 bases each called the read length) into long continuous stretches that are analysed for errors, gene coding regions, and other characteristics.

vi. Finished sequence is submitted to major public sequence databases, making Human Genome Project sequence data thus freely available to anyone around the world (Fig. 18.18).

The human genome reference sequence do not represent any one person's genome. Rather the knowledge obtained is applicable to everyone because all humans share the same basic set of genes and genomic regulatory regions that control the development.

Researchers collected blood (female) or sperm (male) samples from different races like European, African, American (North, Central, South) and Asian ancestry and a few samples were processed as DNA resources.

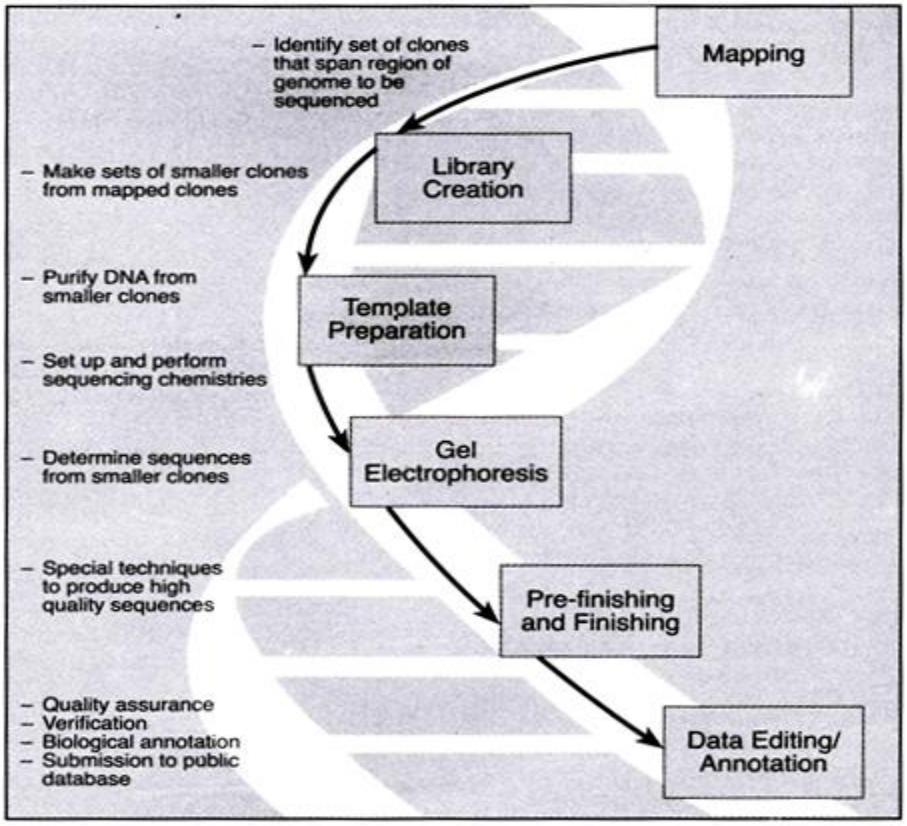


Fig. 18.18: DNA sequencing process

Outcome of Human Genome Project:

i. The human genome contains 3164.7 million chemical nucleotide bases (A, C, T and G).

ii. The average gene consists of 3000 bases, but sizes vary greatly, largest known human gene is **"dystrophin"** – 2.4 million bases.
iii. Total number of genes estimated 30000 approx.

iv. Almost all (99.9%) nucleotide bases are exactly the same in all people.

v. 50% genes are unknown for function.

vi. Less than 2% genomes code for proteins.

vii. Repeated sequences (junk DNA) is 50% of the human genome. This may contribute to create new genes, to modify and reshuffle the existing genes.

viii. A-T rich regions are gene-poor and G-C rich regions are gene-dense. Chromosome-I has the most genes (2968) and the Y chromosome has the fewest (231).

ix. Scientists have identified about 1.4 million locations where single base DNA differences (SNPs) occur in human, these findings will help to localize the disease associated sequences in the chromosomes.

x. Finding the DNA sequences underlying such common diseases as cardiovascular disease, diabetes, arthritis and cancers is being aided by human variation maps (SNPs) generated in HGP.

Announcement of the draft sequence of human genome:

The date 26th June 2000 will be remembered as one of the most important dates in the history of science or even mankind. It was on this day, Francis Collins and Craig Venter, the leaders of the two human genome projects, in the presence of the President of U.S., jointly announced the working drafts of human genome sequence. The detailed results of the teams were later published in February 2001 in scientific journals Nature (IHGSC) and Science (Celera Genomics).

Mapping of the Human Genome:
The most important objective of human genome project was to construct a series of maps for each chromosome. In Fig. 12.1, an outline of the different types of maps is given.

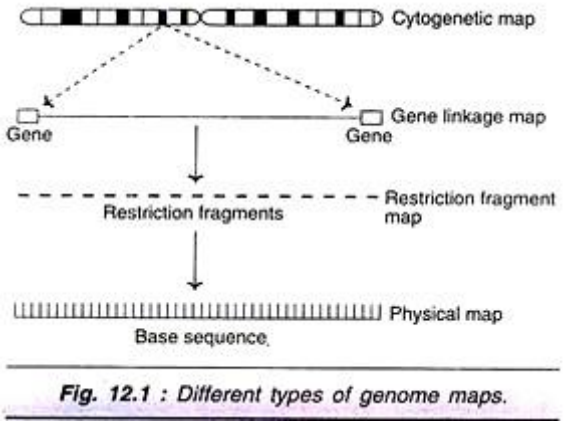


Fig. 12.1 : Different types of genome maps.

1. Cytogenetic map:

This is a map of the chromosome in which the active genes respond to a chemical dye and display themselves as bands on the chromosome.

2. Gene linkage map:

A chromosome map in which the active genes are identified by locating closely associated marker genes. The most commonly used DNA markers are restriction fragment length polymorphism (RFLP), variable number tandems repeats (NTRs) and short tandem repeats (STRs). VNTRs are also called as minisatellites while STRs are microsatellites.

3. Restriction fragment map:

This consists of the random DNA fragments that have been sequenced.

4. Physical map:

This is the ultimate map of the chromosome with highest resolution base sequence. Physical map depicts the location of the active genes and the number of bases between the active genes.

Organization of Human Genome:

An outline of the organization of the human genome is given in Fig. 12.2. Of the 3200 Mb, only a small fraction (48 Mb) represents the actual genes, while the rest is due to gene-related sequences (introns, pseudo genes) and inter-genic DNA (long interspersed nuclear elements, short inter-spread nuclear elements, microsatellites, DNA transposons etc.). Inter-genic DNA represents the parts of the genome that lie between the genes which have no known function. This is appropriately regarded as junk DNA.

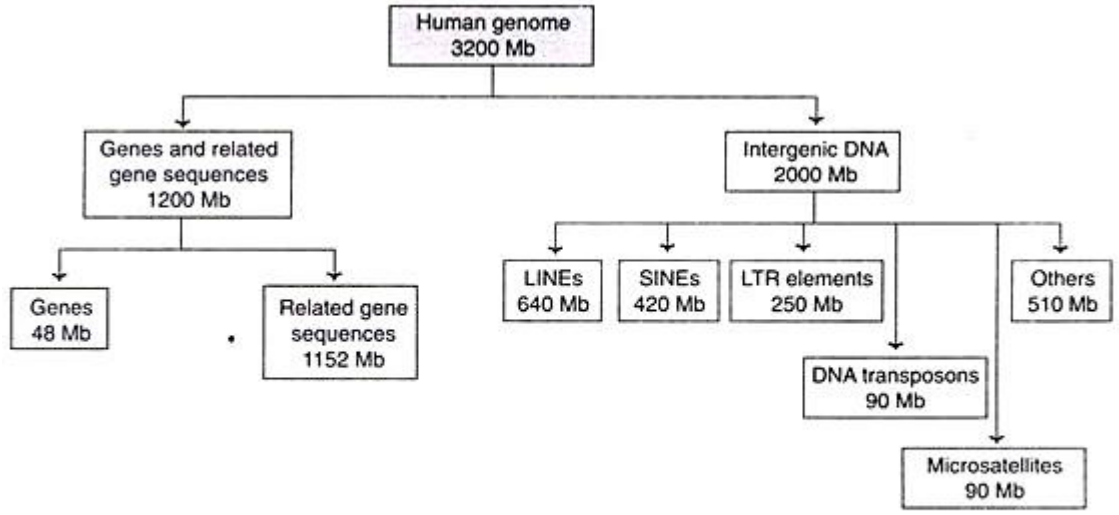


Fig. 12.2 : An overview of the organization of human genome (LINEs-Long interspersed nuclear elements; SINEs–Short interspersed nuclear elements; LTR-Long terminal repeats).

Genes Present in Human Genome:

The two genome projects differ in their estimates of the total number of genes in humans. Their figures are in the range of 30,000-40,000 genes. The main reason for this variation is that it is rather difficult to specifically recognize the DNA sequences which are genes and which are not.

Before the results of the HGP were announced, the best guess of human genes was in the range of 80,000-100,000. This estimate was based on the fact that the number of

proteins in human cells is 80,000-100,000, and thus so many genes expected. The fact that the number of genes is much lower than the proteins suggests that the RNA editing (RNA processing) is widespread, so that a single mRNA may code for more than one protein.

A diagrammatic representation of a typical structure of an average human gene is given in Fig. 12.3. It has exons and introns.

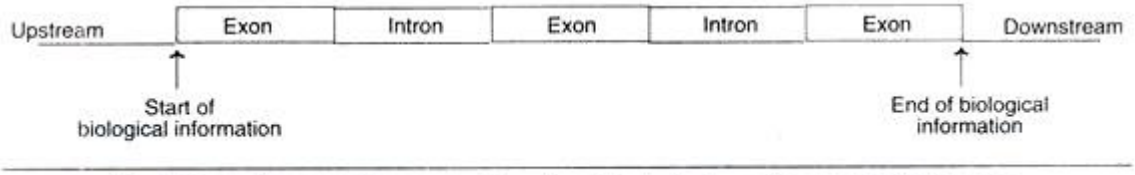


Fig. 12.3 : A diagrammatic representation of a typical structure of an average human gene.

A broad categorization of human gene catalog in the form of a pie chart is depicted in Fig. 12.4. About 17.5% of the genes participate in the general biochemical functions of the cells, 23% in the maintenance of genome, 21% in signal transduction while the remaining 38% are involved in the production of structural proteins, transport proteins, immunoglobins etc.

Human Genes Encoding Proteins:

It is now clear that only 1.1-1.5% of the human genome codes for proteins. Thus, this figure 1.1-1.5% represents exons of genome.

As already described, a huge portion of the genome is composed of introns, and inter-genic sequences (junk DNA).

The major categories of the proteins encoded by human genes are listed in Table 12.4. The function of at least 40% of these proteins are not known.

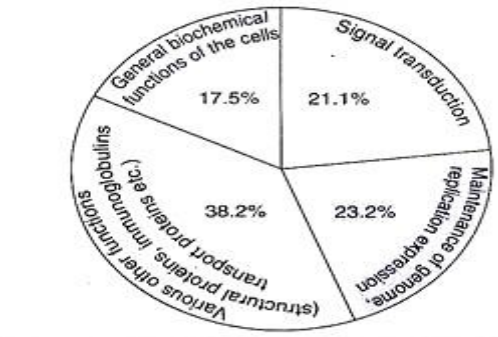


Fig. 12.4 : A pie chart showing a broad categorization of the human gene catalog (About 13000 genes whose functions are not known are not included).

Marked Differences in Individual Chromosomes:

The landscape of human chromosomes varies widely. This includes many features such as gene number per mega base, GC content, density of SNPs and number of transposable elements. For instance, chromosome 19 has the richest gene content (23 genes per mega base) while chromosome 13 and Y chromosome have the least gene content (5 genes per mega base).

Other Interesting/Important Features of Human Genome:

For more interesting features of human genome, refer Table 12.3.

i. It is surprising to note that the number of genes found in humans is only twice that present in the roundworm (19,099) and thrice that of fruit fly (13,001).

ii. Around 200 genes appear to have been derived from bacteria by lateral transfer. Surprisingly, none of these genes are present in non-vertebrate eukaryotes.

iii. The proteins encoded by human genes are more complex than that of invertebrates.

iv. The flood of the data of human genome projects will be highly useful for bioinformatics and biotechnology.

TABLE 12.4 Different categories of proteins encoded by human genes (*based on the Human Genome Project report, 2001*)

Category of proteins	*Percentage*	*Actual number of genes*
Unknown functions	41.0%	12,809
Nucleic acid enzymes	7.5%	2,308
Transcription factors	6.0%	1,850
Receptors	5.0%	1,543
Hydrolases	4.0%	1,227
Regulatory proteins (G-proteins, cell cycle regulators etc.)	3.2%	988
Protooncogenes	2.9%	902
Structural proteins of cytoskeleton	2.8%	876
Kinases	2.8%	868

*(**Note** : This table is based on the rough draft of human genome reported by Celera Genomics. The percentages are derived from a total of 26,383 genes)*

Genomes of Some Other Organisms Sequenced:

Sequencing of genomes is not confined to humans. For obvious reasons and significance, human genome sequencing attracted worldwide attention. In fact, the first genome sequence of the bacteriophage QX174 was determined in 1977. Yeast was the first eukaryotic organism to be sequenced (1986). Recently, the mouse, an animal model closest to human has been sequenced. A selected list of genomes that have been sequenced is given in Table 12.5.

TABLE 12.5 A selected list of genomes that have been sequenced

Name of the species	*Genome size (Mb/kb)*	*Comments (year)*
Bacteriophage QX174	5.38 kb	First genome sequenced (1977).
Plasmid pBR 322	4.3 kb	First plasmid sequenced (1979).
Yeast chromosome III	315 kb	First chromosome sequenced (1992).
Haemophilus influenzae	1.8 Mb	First genome of cellular organism to be sequenced (1995).
Saccharomyces cerevisiae	12 MB	First eukaryotic organism to be sequenced (1996).
Arabidopsis thaliana	125 MB	First plant genome to be sequenced (2000).
Homo sapiens (human)	3200 MB	First mammalian genome to be sequenced (2001).
Oryza sativa (rice)	430 MB	First crop plant genome to be sequenced (2002).
Mus musculis (mouse)	3300 MB	Animal model closest to human (2003).

Ethics and Human Genome:

The research on human genomes will make very sensitive data available that will affect the personal and private lives of individuals. For instance, once it is known that a person carries genes for an incurable disease, what would be the strategy of an insurance company? How will the society treat him/her?

There is a possibility that individuals with substandard genome sequences may be discriminated. Human genome results may also promote racial discrimination categorizing the people with good and bad genome sequences. Considering the gravity of ethics related to a human genome, about 3% of the HGP budget was earmarked for ethical research.

In the 1990s, there was a move by some scientists to patent the genes they discovered. This created an uproar in the public and scientific community. Fortunately, the idea of patenting genes (of human genome sequencing) was dropped. The fear still exists that genetic information will be used for commercial purposes.

Concept of Bioinformatics:

Bioinformatics or life science informatics has emerged as a new branch of biotechnology, offering a fundamental tool to the biologist to accelerate commercialization of biotechnology. Bioinformatics is the classic example of convergence of biotechnology and information technology. Bioinformatics has been the most powerful tools for data mining in life science, analysis, data searching, integration and simulation of molecular biological data.

However, its main priority has been implicated in data storage and genome sequence analysis. The unprecedented growth of information technology and extraordinary growth in molecular biology and recombinant DNA technologies and their interrelated studies culminated into cutting edge technology like bioinformatics. Thus, bioinformatics is also termed as bio-computing or computational biology. Recently, genomics occupies central role in bioinformatics i.e., understanding the basic life process.

Meaning of Bioinformatics:

Bioinformatics is the computer aided study of biology and genetics. In other words, it refers to computer based study of genetics and other biological information. Now the science of bioinformatics is gaining increasing importance in life science especially in the field of molecular biology and plant genetic resources.

Development of Bioinformatics:

The first appeared scientific literature on bioinformatics was published in 1991 by the name “Bioinformatics a new era”. One of the earliest endeavours to build database and create analysis algorithms in prophet, a unix-based software package that allowed scientist to store, analysis and perform mathematical modelling. Infact in 1982, free database called Gen Bank was set up to store DNA sequence data.

This database currently holds about 17 billion bases from more than 1,00,000 genes. In 1980’s, Intell-Genetics developed bioinformatics software called PC/GENE to translate gene sequences to proteins. This programme was meant for predicting protein secondary structure. In 1991, Amos Bairoca introduced the software Swiss-PROT, a protein sequence database.

Currently, Swiss-PROT is a curated protein database under EXPASY (Export Protein Analysis System) proteomics presently with the outcome of remarkable human genome project and making draft sequence available to the people was a landmark in the history of modern biology and science. This has generated immense tool to produce the whole gene catalogues of many microbes and the plant Arabidopsis.

Main points related to bioinformatics are given below:

(i) It is the interface between computer and biology. In other words, it is the application of information technology in the study of biology.

(ii) It utilizes information science for the study of biology.

(iii) It is used for computer based analysis of bio-molecular data especially large scale data set derived from genome sequencing.

(iv) It is used for analysis of data related to genomics, proteornics, metabolomics and other biological aspects.

(v) It has wide applications in handling data related to plant genetic resources.

Branches of Bioinformatics:

The science of bioinformatics can be divided into several branches based on the experimental material used for the study. Bioinformatics is broadly divided into two groups, viz., animal bioinformatics and plant bioinformatics.

Various branches of bioinformatics are defined below:

1. Animal Bioinformatics:

It deals with computer added study of genomics, proteomics and metabolomics in various animal species. It includes study of gene mapping, gene sequencing, animal breeds, animal genetic resources etc. It can be further divided as bioinformatics of mammals reptiles, insects, birds, fishes etc.

2. Plant Bioinformatics:

It deals with computer aided study of plant species. It includes gene mapping, gene sequencing, plant genetic resources, data base etc.

It can be further divided into following branches:

(i) Agricultural Bioinformatics:

It deals with computer based study of various agricultural crop species. It is also referred to as crop bioinformatics.

(ii) Horticultural Bioinformatics:

It refers to computer aided study of horticultural crops, viz. fruit crops, vegetable crops and flower crops.

(iii) Medicinal Plants Bioinformatics:

It deals with computer based study of various medicinal plant species.

(iv) Forest Plant Bioinformatics:

It deals with computer based study of forest plant species.

Computer Programmes used in Biology:

Computers refer to electronic devices which can input, store and manipulate data and output information in a desired form. Now various types of computers such as micro-

computer, minicomputer, mainframe computer, super computer, laptop computer and palmtop computers are available which can be used for multiple purposes.

Various computer programmes are used for the study of biological problems. Such programmes include Microsoft word (MS Word), Microsoft excel (MS excel) and Microsoft power point (MS Power Point).

A brief description of these programmes is presented below:

(i) MS Word:

It is a very useful programme for preparation of project reports, annual reports, writing research papers, varietal information system, plant genetic resources data base, etc.

(ii) MS Excel:

It is useful Computer programme for various types of statistical and biometrical analyses. It can also be used for graphical and diagrammatic display of experimental results.

(iii) MS Power Point:

It is widely used for preparation of slides and presentation of results in various scientific meetings.

Applications of Bioinformatics in Crop Improvement:

Bioinformatics has wide practical applications in genetics and plant breeding.

Some important applications of bioinformatics in plant breeding and genetics are tested below:

1. Varietal Information system

2. Plant Genetic Resources Data Base

3. Studies on Genomics

4. Studies on Proteomics

5. Studies on Metabolomics

6. Studies on Plant Modelling

7. Pedigree Analysis

8. Biometrical Analysis

9. Forecasting Models

1. Varietal Information System:

Bioinformatics has useful applications in developing varietal information system. Variety refers to a genotype which has been released for commercial cultivation (b) State Variety Release Committee or Central Variety Release Committee and notified by the Government of India. Various types of varieties are used in plant breeding.

All such terms are defined below:

TABLE 38.1. Various Types of Varieties

Type of Variety	*Definition/brief description*
1. Primitive Cultivars	Varieties which were selected and cultivated by farmers for many generations, also known as land races.
2. Obsolete Cultivars	Improved varieties of the recent past are known as absolute cultivars.
3. Modern Cultivars	Currently cultivated high yielding varieties are known as modern cultivars.
4. Popular Variety	A widely grown cultivar is referred to as popular variety.
5. Commercial Cultivar	A variety which is used for cultivation on large area.
6. Check Variety	A variety which is used for comparing the performance of stains in breeding experiments is called check variety.
7. Example Variety	In DUS testing, a variety which is used for comparing a particular trait.
8. Reference Variety	All released and notified extant varieties of common knowledge which are under seed multiplication chain. It includes global collection of released varieties.
9. Candidate Variety	A variety which is to be protected under Plant Variety Protection Act is called candidate variety.
10. Extant Varieties	All released and notified varieties which have not been protected under Plant Variety Protection Act are called extant varieties. It includes four types of material, *viz*, Notified varieties, private sector varieties, varieties of common knowledge and farmers varieties.
11. Public Sector Varieties	Varieties developed and released by government organizations are known as public sector varieties.
12. Private Varieties	Varieties developed by private seed companies are called private varieties.
13. Farmers Varieties	Varieties developed by farmers and used for cultivation are known as farmers varieties.
14. Exotic Variety	A foreign variety which is directly recommended for commercial cultivation is called exotic variety.

The detailed information about various type of varieties can be developed using highly heritable characters.

Such information can be used in various ways as given below:

(i) In DUS testing for varietal identification

(ii) In grouping of varieties on the basis of various highly heritable characters.

(iii) In sorting out of cultivars for use in Pre-breeding and traditional breeding.

The information can be stored in the computer memory and be retrieved as and when required.

2. PGB Data Base:

Genetic material of plant which of value as resource for present and future generations of people is referred to as plant genetic resources. It is also known as gene pool, genetic stock and germplasm.

The germplasm is evaluated for several characters such as highly heritable morphological and other characters as given below:

(i) Highly heritable morphological traits

(ii) Yield contributing traits

(iii) Quality characters

(iv) Resistance to biotic and abiotic stresses

(v) Characters of agronomic value.

International Plant Genetic Resources Institute (IPGRI), Rome, Italy has developed descriptors and descriptor states for various crop plants. Such descriptors help in uniform recording of observations on germplasm of crop plants throughout the world. Thus huge data is collected on crop germplasm for several years. Bioinformatics plays an important role in systematic management of this huge data.

Bioinformatics is useful in handling of such data in several ways as follows:

(i) It maintains the data of several locations and several years in a systematic way.

(ii) It permits addition, deletion and updating of information.

(iii) It helps in storage and retrieval of data.

(iv) It also helps in classification of PGR based on various criteria.

(v) It helps m retrieval of data belonging to specific group such as early maturity, late maturity, dwarf types, tall types, resistant to biotic stresses, resistant to abiotic stresses, genotypes with superior quality, genotypes with marker genes, etc.

All such data can be easily managed by computer aided programmes and can be manipulated to get meaningful results.

3. Studies on Genome:

Genome refers to the basic set of chromosome. In a genome each type of chromosome is represented only once. The study of structure and function of entire genome of an organism is referred to as genomics. It is being developed as a sub discipline of genetics which is devoted to the mapping sequencing and functional analysis of genome. The word genomics was coined by Thomas Roderick in 1986.

The discipline of genomics consists of two groups, viz:

(i) Structural genomics and

(ii) Functional genomics.

These are defined below:

(i) Structural Genomics:

It deals with the study of the structure of entire genome of an organism. In other words, it deals with the study of the genetic structure of each chromosome of the basic set of chromosome i.e. genome.

(ii) Functional Genomics:

It deals with the study of genome function. It deals with transcriptome and proteome. Transcriptome refers to complete set of RNAs transcribed from a genome and proteome refers to complete set of proteins encoded by a genome

There are three methods of gene mapping, viz:

(i) Recombination mapping,

(ii) Deletion mapping and

(iii) Molecular mapping.

The last method is widely being used for gene mapping these days. It is computer aided method which is useful in genome mapping. It has been used for genome mapping in various crop plants such as Arabidopsis, rice and maize.

It is a rapid and accurate method of gene mapping. Now computer aided genomic mapping, sequencing and functional analysis studies are being carried out with almost all important field crops. Computer aided programmes have made such studies very simple.

4. Studies on Proteomics:

Proteomics refers to the study of structures and functions of all proteins in an individual. In other words, it deals with the study of entire protein expression in an organism.

Proteomics is of two types, viz:
(i) Structural proteomics and

(ii) Functional proteomics.

These are defined below:

(i) Structural Proteomics:

It refers to the study of the structures of all proteins found in a living organism.

(ii) Functional Proteomics:

It deals with functions of all proteins found in a living organism. In fact, proteomics is a new sub-discipline of functional genomics. It is the study of proteomes which refer to complete set of proteins encoded by a genome. A variety of techniques are used for the study of proteomics. Now computer aided programmes are available for the study of proteomics.

5. Studies on Metabolomics:

Metabolomics refers to the study of all metabolic pathways in a living organism. In other words, it is the computer aided information of all metabolic pathways of a living organism.

Main points related to metabolomics are listed below:

(i) It deals with the study of all metabolic pathways in a living organism.

(ii) It is computer based information about metabolic pathways in a living organism.

(iii) It helps in identification and correction of metabolic disorders in an organism.

(iv) It helps in selection of individuals with normal metabolic pathways.

(v) It helps early detection of genetic disorders associated with metabolic pathways.

6. Modelling of Plants:

Bioinformatics plays an important role in modelling of crop plants. Such computer aided studies have already been made in field pea and several other field crops. First the plant model is conceptualized using various plant traits and then efforts are made to develop such model by using appropriate breeding procedures.

For example, in cotton following characters can be used for developing conceptual plant model:

(i) Maturity duration 160 days

(ii) Plant height 150 cm

(iii) Number of monopodia 2

(iv) Length of sympodia 50 cm

(v) Number of sympodia 20

(vi) Boll weight 4g

(vii) Ginning per cent 38

(viii) Fibre length 28 mm

(ix) Leaf : small and thick

(x) Plant surface—hairy

First donor sources for these traits are identified from the available germplasm. Then efforts are made to combine these traits in one genotype particularly in a popular variety. Such computer based studies help in developing plant ideotype suitable for machine picking and used in multiple cropping system.

7. Pedigree Analysis:

Computer aided studies are useful in pedigree analysis of various cultivars and hybrids. Information about the parentage of cultivars and hybrids is entered into the computer memory which can be retrieved any time. The list of parents that are common in the pedigree of various cultivars and hybrids can be sorted out easily.

It helps in the pedigree analysis which in turn can be used in planning plant breeding programmes especially in the selection of parents for use in hybridization programmes. Through study of protein structures, it helps in pedigree analysis.

8. Biometrical Analysis:

In plant breeding and genetics, various types of biometrical analyses such as correlation, path coefficient, discriminant function, diallel, partial diallel, triallel, quadriallel, generation means, line x tester, triple test cross, stability parameters, D^2 statistics, metroglypth etc. are carried out.

Computer aided programmes are very much useful in carrying out such biometrical analyses. The information obtained from such biometrical analysis is used in better planning of plant breeding programmes for achieving specific goal.

9. Forecasting Models:

Computer aided programmes have wide applications in developing various types of forecasting models especially useful for predicting crop production and productivity and in forecasting incidence of insects and diseases in crop plants. Weather parameters are used in making such predictions. Computer aided remote sensing techniques are used for such predictions.

10. Other Applications:

Besides agricultural applications, bioinformatics have several other useful applications.

Such applications include use of bioinformatics in:

(i) Medical science,

(ii) Forensic science,

(iii) Pharmaceutical and biotech industry.

In medical science computer aided studies are useful in detection of genetic diseases at an early stage of life. It can help in cure of genetic diseases in some cases. The pedigree analysis helps in advising future parents to prevent certain genetic diseases.

In forensic science, bioinformatics is useful in settling disputed cases of children and detecting criminal cases. In pharmaceutical industry, computer aided programmes help in detecting various metabolic pathways involved in the production of a medicine. Thus it can help in mass production of such chemicals.

Advantages of Bioinformatics:

Bioinformatics has several practical applications in genetics and plant breeding as discussed above.

Its main advantages in crop improvement are given below:

1. It provides systematic information about genomics, proteomics and metabolomics of living organisms. This information is useful in planning various breeding and genetical programmes.

2. It helps in finding evolutionary relationship between two species. Studies of nucleotide and protein sequences help in such matter. The closely related organisms have similar sequences and distantly related organisms have dissimilar sequence.

The time of divergence between two species can also be estimated from such studies. Thus bioinformatics helps in the study of evolutionary biology. It helps in drawing phylogenic trees (trees of relatedness).

3. Rapid Method. Is a rapid method of gene mapping and sequencing. Earlier methods of gene mapping were time consuming and pains taking. Bioinformatics has made this task very simple. Now gene hunting has become faster, cheaper and systematic.

4. Identification of similar genes. Computer aided studies help in identification of similar genes in two species. For example, genes similar for biotic and abiotic stresses in two species can be easily detected.

5. High Accuracy. The computer based information has very high level of accuracy and is highly reliable.

6. Bioinformatics has led to advances in understanding basic biological processes which in turn have helped in diagnosis, treatment and prevention of many genetic diseases:

7. It has become possible to reconstruct genes from Expressed Sequence Tags (EST). The EST is nothing but short pieces of genes which can express.

8. Computer aided programmes have made it possible to group proteins into families based on their relatedness.

9. Computer aided programmes are useful in designing primers for PCR. Such primers can be designed with little efforts. Such primers are used to sequence unknown genes or genes of interest.

10. In life science, computer aided programmes are useful in storing, organizing and indexing huge databases.

Limitations of Bioinformatics:

Computer based programmes have helped in better understanding of various processes of life science.

However, there are some limitations of bioinformatics which are listed below:

1. Bioinformatics requires sophisticated laboratory of molecular biology for in-depth study of biomolecules. Establishment of such laboratories requires lot of funds.

2. Computer based study of life science requires some training about various computer programmes applicable for the study of different processes of life science. Thus special training is required for handling of computer based biological data.

3. There should be uninterrupted electricity (power) supply for computer aided biological investigations. Interruption of power may sometimes lead to loss of huge data from the computer memory.

4. There should be regular checking of computer viruses because viruses may pose several problems such as deletion of data and corruption of the programmes.

5. The maintenance and up keeping of molecular laboratories involves lot of expenditure which sometimes becomes a limiting factor for computer based molecular studies.

Probable Questions:

1. Define genome and genomics.

2. Write down the types of genomics.

3. What do you mean by functional genomics? How functional genomics is studied?

4. How Human genome project was evolved.

5. What are the goals of HGP?

6. Write down the main features of human genome.

7. Write down the outcomes of HGP.

8. Write down different types of mapping.

9. What are limitations of Bioinformatics?

10. What are the advantages of bioinformatics?

11. Discuss different branches of bioinformatics.

12. Discuss different branches of genomics and proteomics.

13. Define bioinformatics. Which computer programmes are used in bioinformatics study?

Suggested Readings:

1. Molecular Cell Biology by Lodish, Fourth Edition.
2. The Cell – A Molecular Approach by Cooper and Hausman, Fourth Edition
3. Principles of Genetics by Snustad and Simmons, Sixth Edition.
4. Molecular Biology of the Cell – by Bruce Alberts
5. Cell and Molecular Biology by Gerald Karp, 7th Edition.
6. Gene cloning and DNA Analysis by T. A. Brown, Sixth Edition.
7. Genetics . Verma and Agarwal.

Disclaimer:

The study materials of this book have been collected from various books, e-books, journals and other e sources.