Chapter 5. Genomics, Proteomics and Bioinformatics

5.1 Genomics:

Genomics is a relatively new discipline. Although, the DNA was first isolated as early as 1869, it took more than one Century for the first genomes to be sequenced. The term genomics was introduced recently by Thomas Roderick in 1986. Genomics describe the detailed study of genome; it is structural organisation and function using various modern methods including computational biology. It involves the genome sequencing and computer aided analysis to understand its structural organisation and functions, genome mapping and related studies. The term genome represents the complete genetic material including both nuclear and cytoplasmic genes present in a cell. The Human Genome Project (HGP), sponsored in the United States by the Department of Energy and the National Institutes of Health, has created the field of genomics understanding genetic material on a large scale. The field of molecular life science is changing rapidly, because of the genomic revolution. Revolutionary improvements in the DNA sequencing techniques have given rise to a large amount of DNA sequences, which is difficult to manage, particularly for future references and analysis. Technological developments in computer and information technology have helped a lot in managing the huge data of DNA sequences in the form of computerised databases and it is access through internet.

5.1.1 Concept of genomics:

Thormas Roderich introduced the term genomics in 1986. It is scientific method of mapping, sequencing and analysing and making the use of genetic information for further use in multifarious area. Genomics can be defined as the study of molecular organisation of genomes, their information contents and the gene products they encode.

"Genomics is the study of structure and functions of a genome of an organism. It concerns with the sequencing and analysis of an organism's genome. The genome is nothing but the total DNA content that present within one cell of an organism".

5.1.2 Types of genomics:

In the last few years, some interesting findings have been recorded and several new branches have emerged. Consequently, the area of genomics has quietly widened. However, the genomics is broadly categorised into three types namely, structural genomics, functional genomics and comparative genomics.

1) Structural Genomics: The process of finding out the sequences of genome is called as structural genomics. The structural genomics deals with DNA sequencing, sequence assembly, sequence organisation and management. Structural genomics attempts to determine the structure of every protein encoded by the genome, rather than focusing on one particular protein. Basically, it is the starting stage of genome analysis i.e. construction of genetic map or sequence maps of high resolution of the organism. The complete DNA sequence of an organism is its ultimate physical map. Due to rapid advancement in DNA technology and completion of several genome sequencing projects for the last few years, the concept of structural genomics has come to a stage of transition. Now structural genomics also includes systematic and determination of 3D structure of proteins found in living cells, because proteins in every group of individuals vary and so there would also be variations in genome sequences.

2) Functional Genomics: To study and understand the function of gene is the basis of functional genomics. Based on the structural genomics the reconstruction of genome sequences is useful to find out the function that the genes do. It gives an idea of function of all gene sequence and their expression in organism. The different tools useful for structural genomics are bioinformatics sequences, DNA chips, 2D gels etc. This information lends support to design experiment to find out the functions that specific genome does. The strategy of functional genomics has widened the scope of biological investigations. This strategy is based on systematic study of single gene or protein to all genes. Therefore, the large-scale experimental methodologies characterise the functional genomics. Hence, the functional genomics provide the novel information about the genome. This eases the understanding of genes and function of proteins and protein interactions. The development of microarray technology and proteomics helped to explore the instantaneous events of all the genes expressed in a cell or tissue present at varying environmental conditions like temperature, pH, etc.

3) Comparative Genomics: The complete genome sequences of cellular organisms become available, the notable finding was recorded. It was found that one third of the genes encoded on each genome had no predictable or known function. e.g. in *E.coli* K_{12} about 40 % genes have unknown function. The level of evolutionary conservation of microbial proteins is rather uniform with about 70 % of gene products from each of sequenced genomes having homologous in distinct genomes. The function of these gene can be predicted by comparing different genomes and by transferring functional annotations of protein for better studies organisms to their orthologs (the same gene in different species that connect) as opposed to paralogs i.e., genes related by duplication within the genome from less studied organism. For

better understanding of genomes, this makes comparative genomics as a powerful approach. Comparative genomics includes several aspects such as analysis of protein sets from completely sequenced genomes. General purpose databases and organisms specific databases used for comparative genomics.

5.1.3 Methods used for whole genome sequencing

The genome, of an organism (bacteria, virus, potato, human) is made up of DNA. Each organism has a unique DNA sequence which is composed of bases (A, T, C, and G). If the sequence of the bases in an organism are known, then we can identify its unique DNA fingerprint, or pattern. Determining the order of bases is called sequencing. Whole genome sequencing is a laboratory procedure that determines the order of bases in the genome of an organism in one process.

There are several methods used for whole genome sequencing. Sequencing of genome chiefly comprises three steps: i) the cloning of the DNA to be sequenced, ii) the sequencing reactions and electrophoretic separations and iii) the analysis of ensuing data. Following are important methods of whole genome sequencing:

1) Chemical Methods:

This method was developed by Maxam and Gilbert (1977). A restriction fragment of DNA is labelled with 32p at either its 5' or 3' using either of the enzymes polynucleotide kinase or terminal transferase. From a restriction map, an enzyme is selected to remove a small piece from one end of the molecule leaving just one end labelled. The DNA is then chemically cleaved at specific residues in five different reactions. These reactions are partially completed and partial digestion products are separated on a polyacrylamide gel and autoradiographed. The fragments having the labelled terminus are seen.

2) Whole Genome Shotgun Sequencing:

J. Craig Venter and H. Smith developed whole Genome shotgun sequencing and the two genome of bacteria *Haemophilus influenzae* and *Mycoplasma genitalium*. This method consists of four steps:

i) Library Construction: The chromosome is isolated from the desired cells following the methods of molecular biology. The isolated DNA is randomly fragmented into small pieces using ultrasonic waves. Then fragments are purified and attached to plasmid vectors. Plasmid with single insert is isolated. A library of plasmid clones are prepared by transforming *E. coli* strains with plasmid that lacked restriction enzymes.

ii) Random Sequencing: The DNA is purified from plasmid. Thousands of DNA fragments are sequenced using automated sequencer by using primers labelled with special dyes. Normally with universal primers, thousands of templates were used. These recognise the plasmid DNA sequences next to bacterial DNA insert. The whole genome is sequenced several times.

iii) Fragment - alignments and Gap Closure: By using special computer programme, the sequenced DNA fragments are clustered and assembled into longer stretches of sequence by comparing nucleotide sequence overlaps between fragments. Two fragments are joined to form a large stretch of DNA if the sequences at their ends overlapped and matched. This overlap comparison method resulted in a set of larger contiguous nucleotide sequence called contigs. The contigs are aligned in a proper order to form the completed genome sequence.

iv) Proof Reading: Then the proof reading of sequences is done carefully so that any ambiguities in the sequence could be resolved. The sequence is also checked for the presence of any frame shift mutation; if so, the mutation is corrected.

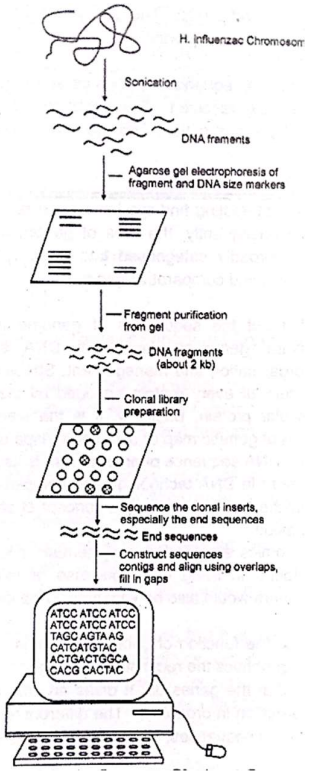


Fig. 5.1 : Whole Genome shotgun sequencing

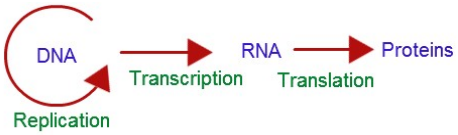
5.2 Proteomics:

Proteomics is the study of all the proteins produced by a cell. Proteomics is the identification, analysis and large scale characterisation of proteome expressed by any cells, tissues and organs under the defined conditions. The major objectives of proteomics are: i) to

characterise post-transcriptional modifications in protein and ii) to prepare 3D map of a cell indicating the exact location of protein.

5.2.1Concept of proteomics:

The total protein component in a cell or organism is referred as the proteome. Proteomics deals with the study of proteomes. In broader term, proteomics is defined as the total protein content of a cell or that of an organism. The terms 'proteome' and 'proteomics' were coined in the early 1990 by Marc Wilkins. Proteomics helps in understanding of alteration in protein expression during different stages of life cycle or under stress condition. Likewise, Proteomics helps in understanding the structure and function of different proteins as well as protein - protein interactions of an organism. A minor defect in protein structure, it is function or alternation in expression pattern can be easily detected using proteomics studies. This is important with regards to drug development and understanding various biological processes, as proteins are the most favourable targets for various drugs. The first protein studies that can be called proteomics began in 1975 with the introduction of the two dimensional gel and mapping of the proteins from the bacterium Escherichia coli, guinea pig and mouse. Proteins are macromolecules; long chains of amino acids. This amino acid chain is constructed when the cellular machinery of the ribosome translates RNA transcripts from DNA in the cell's nucleus. The transfer of information within cells commonly follows this path from DNA to RNA to protein.



5.2.2 Types of proteomics:

The types of proteomics are as follows:

1) **Structural Proteomics:**

Structural proteomics deals with the study of structure and nature of protein complexes present in a particular cell organelle. It is mapping out the 3-D structure and nature of protein complexes present specifically in a particular cell organelle. The ultimate aim of structural proteomics is to build a body of structural information that will help predict the probable structure and potential function for almost any protein from knowledge of its coding sequence. Structural proteomics can also help assembling information about protein - protein interactions and about architecture of cells to explain how the expression of certain proteins contributes in cell's unique characteristics.

2) Functional Proteomics:

Functional proteomics refers to the use of proteomics techniques to analyse the characteristics of molecular protein-networks involved in a living cell. One of the recent successes of functional proteomics is the identification and analysis of molecular protein networks involved in the nuclear pore complex (NPC) in yeast. This success helps understand the translocation of molecules from nucleus to the cytoplasm and vice versa.

3) Expression Proteomics:

Expression proteomics concerned with to the quantitative study of protein expression between samples differing by some variable. The pattern of expression of the complete proteome or of its part (sub-proteome) between samples can be compared with the help of expression proteomics. The expression proteomics is quite useful in identifying disease specific proteins. For example, over expression or under-expression of proteins in cancerous cells and normal cells taken from a cancer patient and a normal individual, respectively, can be analysed using various techniques, such as two dimensional gel elec trophoresis, mass spectrometry, microarray, etc. This can help understand the development of cancer and facilitate development of drugs to treat cancer.

5.2.3 Methods used in proteome analysis:

Although new methods in proteomics are being developed, the traditional methods are; two-dimensional electrophoresis, and mass spectrometry. The first dimension uses iso-electric focusing and second dimension is SDS-PAGE. Some of the methods used in proteome (protein) analysis are as follow:

a) Sodium Dodecyl Sulphate Polyacrylamide Gel Electrophoresis (SDS-PAGE):

Separation of some of the proteins dose not occur due to similar charge: mass ratio. Therefore, such proteins are treated first with an ionic detergent called sodium dodecyl sulphate (SDS) before to electrophoresis (PAGE). Therefore, such electrophoresis is called SDS-PAGE electrophoresis.

SDS-PAGE is a high resolution method used universally for analysing the mixture of proteins according to their respective size. SDS solubilised in soluble protein makes possible the analysis of the other insoluble mixture. Separation of the proteins doses not occur due to similar charge: mass ratio (z/m). Therefore, such proteins are treated first with an ionic detergents SDS before the start and during the course of electrophoresis.

Identical proteins are denatured by SDS resulting in their sub-units. The polypeptide chains get opened and extended. On the basis of their mass but not the charge, the molecules are separated. Electrophoretic separation is normally used for these reasons i.e. (i) gel acts as molecular sieves hence separates the molecules the molecules on the basis of their size, and (ii) gel suppresses conventional currents produced by small temperature gradient which improve the resolution. Polyacrylamide gel is used for this purpose due to its good nature (chemically inert, stable over a wide range of pH, temperature and transparent). Polyacrylamide gel is better for size fraction of proteins.

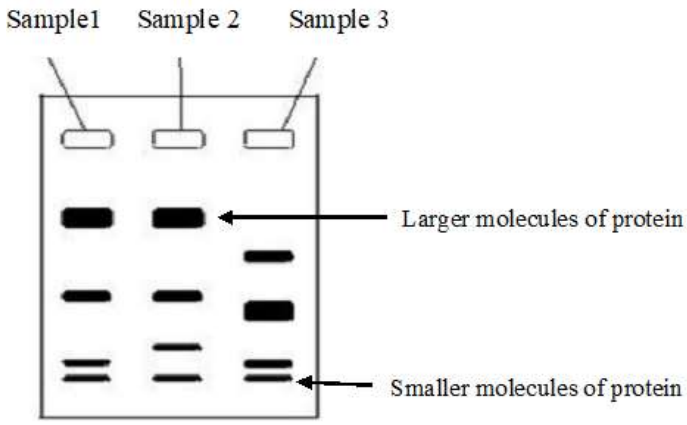


Fig. 5.2 SDS-PAGE analysis of proteins

The proteins are denatured and have negative charge with a uniform charge to mass ratio(z/m) when treated with SDS. Proteins migrates towards anode at alkaline pH through PAGE gel during electrophoresis. The smaller polypeptides moves faster followed by the larger polypeptides. Therefore, intrinsic charge on proteins is masked in SDS-PAGE. Hence separation is based on size. Molecular weight of the separated protein can be analysed by comparing the molecular weight of the standard protein and its mobility. In analysis of a complex a complex mixture of proteins the resolution is improved by the initial movement through a stacking gel. The final bands in the separating gel are sharper and focused in better way.

Two dimensional electrophoresis is very useful and effective method as it separating proteins and can resolve thousands of proteins in a mixture.

b) Iso-electric Focusing (IEF):

The biomolecule like proteins have electric charge which depends on molecule to molecule and conditions of medium (pH of buffer in which dissolved). Charged molecules can

be separated by electrophoresis in gels. Due to the differences in amino acid composition proteins have net charge or iso-electric points (no charge) as a given pH of buffer.

The atmospheric substances such as proteins which differ in their isoelectric points can be separated by IEF. Isoelectric point is a pH value at which the net charges on molecules are zero. Ampholytes (i. e. complex mixture of synthetic polyamino-polycarboxylic acids) are introduced into gel to create the pH gradient (wide range from 3 to 10, or narrow range of 7 to 8). Then potential difference is applied across the gel. The molecule having difference in isoelectric points by a little as 0.01 pH unit can be separated. Proteins migrate depending on their charge until they reach a region which pH corresponds to respective iso-electric points at which pH proteins possess no net charge and hence got focused.

c) Mass spectrometry:

Mass spectrometer which employed fixed magnetic and electric field to separate ions of different mass and energy. Two-dimensional electrophoresis is more powerful when coupled with mass spectrometry. The unknown protein spot is cut from gel and cleaved by trypsin digestion into fragments which are then analysed by mass spectrometer and mass of fragments is plotted. This mass finger print can be used to estimate the probable amino acid composition of each fragment and tentatively identify the protein. The proteome and its charges can be studied very effectively by employing the two techniques together.

Mass spectrometry can also provide valuable information about covalent modification of proteins which can affect their activity. Mass spectrometry is very useful technique. It is used in identification of unknown compounds and determination of structural and chemical properties of compounds when present in small amount (10^{-4} -10^{-8} g). This technique involves: (i) the production of ions of the material in sample, (ii) their separation on the basis of their mass change (m:e), and (iii) determination of relative abundance of each ion.

Therefore, mass spectrometer consists of three components: the source of ion, an analyser, and a detector. It dose not directly measure the molecular mass but detects m:e ratio. Mass is measured in terms of Dalton (Da). One Dalton = $1/12^{th}$ mass of a single atom of isotonic carbon (^{13} C).

In recent days, mass spectrometry has become an essential tool for analysis of genome and proteome in its many forms. It is capable of identifying and characterising proteins present even in picomoles (10^{-12}).

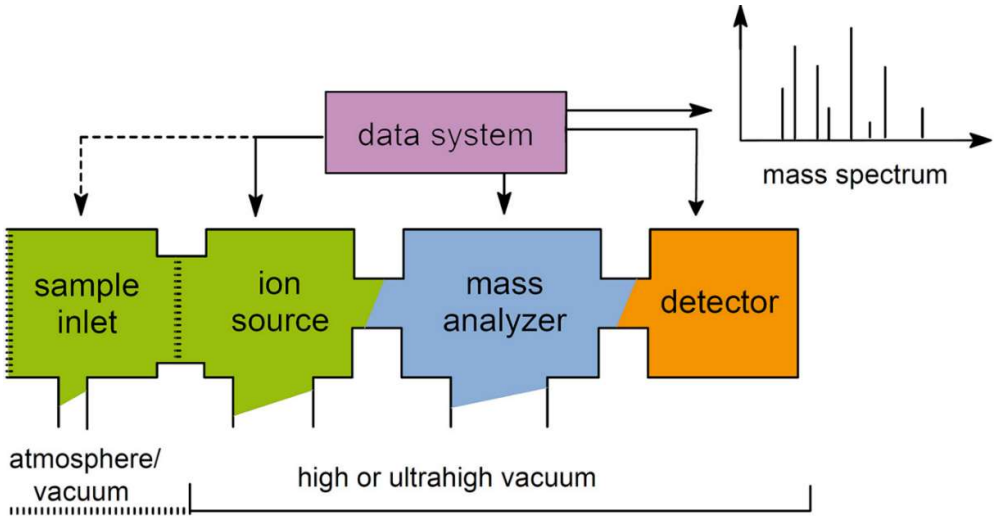


Fig. 5.3: Flow diagram of mass spectrometry

5.3 Bioinformatics:

Term 'Bioinformatics' was coined by Paulien Hogeweg and Ben Hesper in 1970. Basically, bioinformatics deals with the information in the fields of information Technology, Computer Science and Biology. Biologist performs research in laboratory and collects information about DNA and protein sequences, gene expressions etc. Computer scientists are involved in developing algorithms, tools, software to store and analyse data. Bio informaticians study biological questions by analysing molecular data with various programs and tools. Today, bioinformatics is used in number of fields such as microbial genome applications, biotechnology, gene therapy, agriculture etc. The term bioinformatics has been derived by combining biology and informatics. Bioinformatics is the field of science in which biology, computer science and information technology merges to form a single discipline. It is the emerging field that deals with the application of computers to the collection, organisation, analysis, manipulation and presentation and sharing of biologic data to solve biological problems on the molecular level.

5.3.1 Concept bioinformatics:

"Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics and engineering to analyse and interpret biological data."

Bioinformatics is the application of computer technology to the management of biological information. Computers are used to collect, store, analyse and integrate biological

and genetic information which can then be applied to gene-based drug discovery and development. In another words, bioinformatics is the application of information sciences (mathematics, statistics and computer sciences) to increase our understanding of biology, biochemistry and biological data. In broad sense, bioinformatics can be considered as information technology applied to management and analysis of biological dada. Bioinformatics has been used for in silico analyses of biological queries using mathematical and statistical techniques.

Bioinformatics is the mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information. Thus, bioinformatics is a multidisciplinary science which aims to use the benefits of computer technologies in understanding the biology of life. In short, bioinformatics is the management and analysis of biological information stored in database.

5.3.2 Database and its classification:

Database:

A database is a repository of sequences (DNA or amino acids) which provide a centralised and homogeneous view of its contents. The repository sequences is created and modified through a database management system (DBMS). Every data item in the database is structured according to a scheme, defined as a set of prespecified rules through the data definition language. The information is stored in database, many of which are accessible to everyone on the internet. The contents of database can be accessed through a graphical user interface (GUI) that allows browsing through the contents of the repository very much similar as one may browse through the books in library.

Classification of Databases:

The databases are mainly classified into two categories viz:

a) Sequence Databases: This database involves the sequences of both proteins and nucleic acids.

b) Structural Databases: This database involves only sequences of proteins. In addition to this database is also classified into following three categories:

i) Primary Databases: It contains information of the sequence or structure alone of either protein or nucleic acid. e.g. PIR or protein sequence, Genbank and DDBJ for genome sequence. Primarydatabase tools are effective for identifying the similarities, but analysis of

output is sometimes difficult. GenBank obtained more than a million of sequences from more than 18,000 organisms.

ii) Secondary Databases: The secondary databases contain derived-information from the primary databases. e.g. information on conserved sequence, signature sequence and active site residues of protein families by using SCOP. It is more useful than the primary databases. Orthology provides an important layer of information when considering phylogenetic relationships between the genes. Depending on the type of analysis method used, relationship may be elucidated in considerable detail including superfamily, family, subfamily and species specific sequence levels.

iii) Composite Databases: It is obviating the need to search multiple resources. The SCOP is structural classification of proteins in which the proteins are classified into hierarchial levels such as classes, folds, super-families. A modern database pertaining to protein sequence and structural correlations on the 'NET' was established by Bairoch (1991). This database was called PROSIT which later on was strengthened with a database on sequence analysis and comparison of protein sequence known as SEQUANALREE.

5.3.3 Data retrieval tools:

Gene Bank contains 7 millions sequence record covering about 9 million nucleotide bases. Unless the databases are easily searched and entries retrieved in a usable and meaningful format, biological databases serve a little purpose. Moreover, efforts made on sequencing will not be meaningful if biological community as a whole cannot make use of the information hidden within millions of bases and amino acids. There are several database retrieval tools such as ENTREZ, OMIM, BLAST, FASTA etc.

1) ENTREZ:

The integrated information database retrieval system of NCBI is called Entrez. It is most utilised of all biological database system. Entrez system is useful to accesses nucleotide and protein sequence data and get structural data (3D), genome map. Protein structures from MMDB (Molecular Modelling Database) and the biomedical literature via PubMed, with embedded links to the NCBI taxonomy. The sequences in Entrez, especially protein sequences are obtained from a variety of database sources such as gene bank protein translations, protein identification resource, SWISS-PROT, Protein Research Foundation, Protein Data Bank and reference sequences and therefore include more sequence data than GenBank alone.

Entrez is not a database, but it is the interface through which all of its component databases can be accessed and traversed. Entrez has ability to retrieve the related sequence

structures and references. Entrez provides text searching of sequence or bibliographic records and extensive links to related information. Some links are simple cross-references, for example, from a sequence to the abstract of the paper in which it was reported, from a protein sequence and its corresponding DNA sequence or to alignments with other sequences. The resulting pre-computed document 'neighbours' allow rapid access for browsing groups of related records. A project called LinkOut was recently implemented to expand the range of external links from individual database records to relate outside services, including organism-specific genome databases.

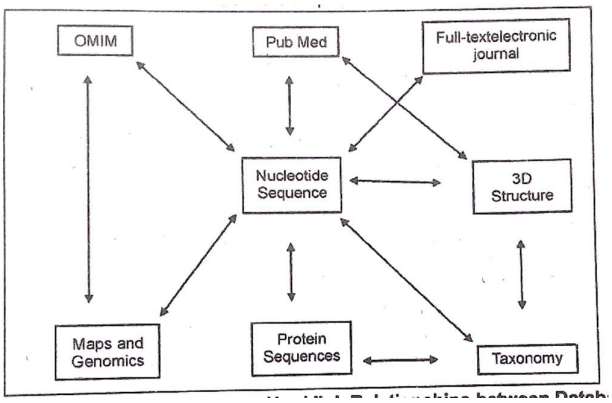


Fig. 5.2 Entrez Map

The Entrez information includes PubMed records, nucleotide and protein sequence data, 3D structure, information and mapping. All the information can be accessed by issuing only one query.

2) OMIM:

OMIM (Online Mendelian Inheritance in Man) is a non-sequence information resource that is very much useful in genomics. It is web based electronic version catalogue that contains thousands of entries for human genes and genetic disorder. It serves as a phenotypic companion to Human Genomic Project. It was founded by Victor McKusick at the johns Hopkins University. Concise textual information is provided by OMIM from the published literature on the conditions of human having genetic disorders and full citation information. At the NCBI, the outline version of OMIM is housed. In addition to this, links are provided to Entrez from all references cited within each OMOM entry. Internet resource for OMIM is: http://www.ncbi.nlm.nih.gov/omim.

The OMIM cytogenetics and morbid maps present cytogenetics locations for those genes with published locations and provide an alphabetical list of all the diseases described in OMIM. Therefore, it is necessary to consider the results of web based biology reported in the

scientific literature in order to validate the findings generated through computer based comparative analysis. Hence, integration of scientific data with the literature is an important step for creating a unified information resource in life science. For this purpose, individuals are provided with a direct link from OMIM to PubMed, the NCBI literature system.

It is simple and very easy to perform OMIM searches. A simple query is performed by search engine on the basis one or more words typed into a search window. Consequently, a list of documents is returned containing the query words. The users can select one or more disorders from the list so as to see the full text of OMIM entry.

3) BLAST:

Sequenced data are compared to one another using the Basic Local Alignment Search Tool (BLAST). BLAST is a programme for sequence similarity searching developed at the NCBI. It identifies genes and genetic features. It is BLAST that provides a method for rapid searching of nucleotide and protein databases. It executes sequences searches against the entire DNA database in less than 15 seconds. This search tool is better for proteins than for nucleotides.

BLAST information guide is designed to assist new users in employing NCBI tools such as BLAST and PSI-BLAST in their research. Sequence alignments provide a powerful way to compare novel sequences with previously characterised genes. Both functional and evolutionary information can be inferred from well-designed queries and alignments. Since, the BLAST algorithm detects local as well as global alignments, regions of similarity embedded in otherwise unrelated proteins could be detected. Both types of similarity may provide important clues to the function of uncharacterised proteins. There are several variants of BLAST each is distinguished by types of sequences (DNA or protein) of the query and database sequences. These are as follow:

i) BLASTp: Blastp compares an amino acid query sequence against a protein sequence database.

ii) BLASTm: It compares a nucleotide query sequence against a nucleotide sequence database.

iii) BLASTx: Blastx compares a nucleotide query sequence translated in all reading Frames against a protein sequence database.

iv) TBLASTn: tblastn compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames.

v) tblastx: tblastx compares the six-Frame translations of a nucleotide query sequence against the six-Frame translations of a nucleotide sequence database (http:// www. ebi.ac. uk/ services/).

4) FASTA:

FASTA was the first widely used programme for database similarity search. It is sequence comparison software that uses the method of Pearson and Lipman. FASTA performed optimised search for local alignment using a substitution matrix. It takes sufficient time to apply this strategy thoroughly. This programme uses the observed pattern of word hits to identify potential matches before attempting the more time consuming optimised search. The Ktup parameter controls the trade off between speed and sensitivity.

FASTA format contains a definition line and sequence characters. It may be used as input of many analysis programmes. FASTA format is used in a variety of molecular software suites. BLAST is faster and sensitive than FASTA in detecting more alignments, but FASTA returns fever false hits.

The basic FASTA algorithm assumes a query sequence and a database over the same alphabet. It searches a DNA sequence in a DNA database or a protein sequence in a protein database. Practically, FASTA is a family of programme, allowing also queries of DNA versus a protein database or vice versa. In these variants, there is further distinction, which regards the location of gaps; one may assume that gaps occur only in the codon frames corresponding to amino acid insertion; alternatively, one can assume gap location to be arbitrary accounting for insertion or deletion of nucleotides. This search tool is preferred for searching nucleotides.

EXERCISES

1. Give the concept of genomics.
2. Give the concept of proteomics.
3. Give the concept of bioinformatics.
4. Enlist the types of genomics.
5. Describe in detail the methods used for whole genome sequencing.
6. Describe the different types of genomics.
7. Describe the different types of proteomics.
8. Describe the methods used in proteome analysis.
9. What is database?
10. Describe the classification of database.
11. Describe data retrieval tools.

12. Write short note on ENTREZ.
13. Write short note on OMIM.
14. Write short note on BLAST.
15. Write short note on FASTA.