WHAT IS BIOINFORMATICS?

Bioinformatics is a newly emerged scientific discipline for the computational "analysis and storage of biological data". It is derived from two words: Bio means biology; Informatique (a French word) meaning 'data processing'. The term "bioinformatics" was coined by Paulien Hogeweg and Ben Hesper in 1978 for the study of information processes in biological systems. It is the field in which biology, computer science and information technology merge into single discipline for managing and analyzing biological data using advanced computing techniques. It has emerged as a full-fledged inter-disciplinary subject that interfaces the developments of computer science and information technology with biological sciences. The knowledge of computer science and information technology is applied for creation as well as management of databases, data warehousing, data mining and overall communication networking throughout the world.

Bioinformatics can be defined as the storage, analysis, and searching retrieval of data (e.g. nucleic acid sequences for the genes and RNAs, amino acid sequence and structural information of protein). It is defined as "bioinformatics more precisely as the mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information" (Fredj Tekaia). However, according to Richard Durbin, "all biological computing is not bioinformatics, e.g. mathematical modeling is not bioinformatics, if the model is related to biological problems." In his opinion, bioinformatics has to do with management and the subsequent use of biological information, particularly the genetic information.

Bioinformatics is "Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive. analyze, or visualize such data" (National Institute of Mental Health, Bathesda (USA), 2000.)

Bioinformatics particularly focuses on subdomains of biology by developing biological databases and algorithms which helps in the research on genes and proteins. It fetched much attention during the Human Genome Project. This had enabled to sequence the large and complex human genome, and to identify the genes in human DNA. The researchers working in this field to process and analyze the generated data are using powerful computers and special algorithms.

AIMS OF BIOININFORMATICS

Due to the spectacular growth of Biotechnology and Molecular Biology tremendous amount of data on nucleotide sequences are being produced. The main aims of bioinformatics are,

(1) To uncover the wealth of biological information hidden in the mass of nucleotide sequence.
(2) To know the amino acid sequence on the basis of nucleotide sequences.
(3) To know the structure of proteins on the basis of amino acid sequences.
(4) Prediction of functional aspects of proteins on the basis of its structure.
(5) To provide biological data information and other related literature on the Internet.
(6) To obtain a clearer insight into the fundamental biology of organisms.
(7) Using this information for welfare of mankind.

Therefore, it is clear that the knowledge of bioinformatics not merely limited to the computation of data, but in reality it can be used to solve many biological problems and can be applied how living things work. The major applications of bioinformatics is to access, search, visualize and retrieve the information of databases of the sequences as well as to understand structural information of biomolecules protein analysis etc. Other applications include cell metabolism, biodiversity, downstream processing in chemical engineering, drug and vaccine design. Current efforts in molecular biology (e.g. genome projects) are producing a large quantity of data that is not only providing exciting opportunities for knowledge discovery, but also increasing problem of information overload. Bioinformatics also concerns the development of new tools for the analysis of genomic and molecular biological data. This can be applied to all fields of biological science as well as agricultural science, environmental science, pharmaceutical science, chemical science and medical science.

SCOPE AND RESEARCH AREAS OF BIOININFORMATICS

The various applications of bioinformatics include:

SEQUENCE ANALYSIS

The most recognized application of bioinformatics is sequence analysis. In this application, DNA of various organisms are sequenced and stored as databases for easy retrieval and comparison.

A. NUCLEOTIDE SEQUENCE ANALYSIS

Since the sequence of Phage ΦX174 in l977, hundreds of DNA sequences of organisms have been decoded and their nucleotide sequence has been stored in databases. As amount of data of nucleotide sequences has enormous growth, it became impossible to analyze it manually since it contains billions of nucleotides. The information is analyzed by utilizing bioinformatics tools to determine regulatory sequences as well as coding sequences. A comparison of genomes within a species or between different species can be done to find out similarities and differences.

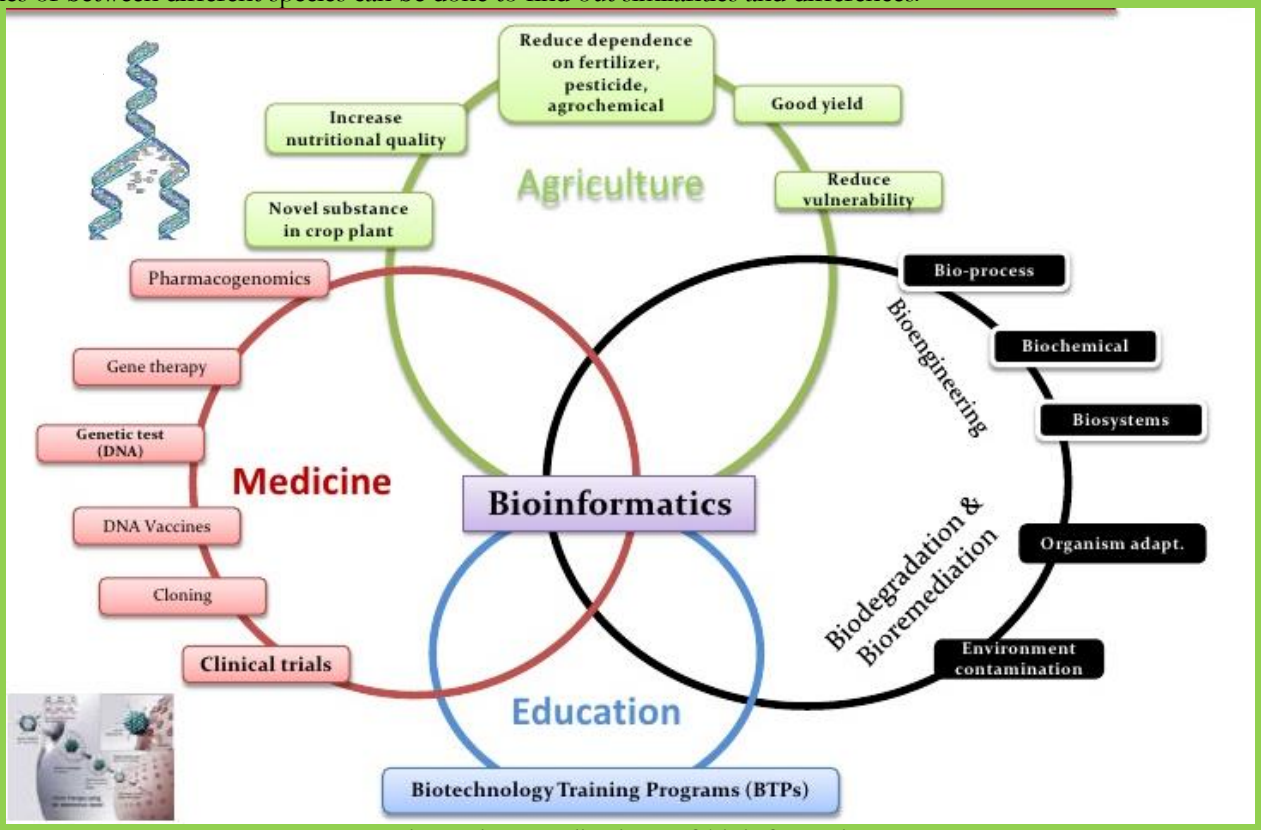

Fig: Various applications of bioinformatics

One of the most important application of the bioinformatics is alignment of two sequences to determine
(l) What parts of the sequences are conserved from one species to the next?
(2) Flow much an organism has diverged from other organisms, which is done simply by comparing their DNA sequences? More similar two gene sequences are to one another, more closely the organisms are related. On the contrary, dissimilarity of the two sequences shows the distance between the two genes in relation to each other. With this application we can compare sequences to determine how organisms have diverged during evolution.

B. SEQUENCE TRANSLATION

lf we know the nucleotide sequence, then it can be converted into the amino acid sequence (translation) or vice versa. i.e., protein sequences into nucleotide sequences or complimentary DNA (cDNA) (back translation). Molecular biologists need to analyze their nucleotide sequence, and the best way to do this is to study the protein product.

C. PROTEIN SEQUENCE ANALYSIS

After obtaining the amino acid sequence of a protein, there are numerous valuable tools that allow further analysis molecular weight (MW), isoelectric point (pI), titration curves, hydrophobicity etc. of the proteins from the sequence. Another important tool is sequence alignment applications that helps to find degree of similarity between two or several proteins.

Numerous databases exist and each database is accessible through convenient search programs. The main databases throughout the globe containing information useful for computational biologists include the National Center for Biotechnology Information (NCBI). the European Bioinformatics Institute (EBI), and the DNA Data Bank of Japan

(DDBJ). These can be used for sequence retrieval (to find the nucleotide sequence for a gene of interest) and sequence identification (to find function and possible origin of gene from a sequence).

D. STRUCTURE ANALYSIS

Several programs have been developed to visualize the three dimensional shape of proteins and nucleotides that furnish an excellent three dimensional view on the computer monitor. Observing a protein in 3-D gives greater understanding of protein structure and function.

PHARMACOINFORMATICS

The aspects of bioinformatics that can be applied to drug designing are known as Pharmacoinformatics. It is also known as Biomedical Informatics or the Medical informatics. It is an emerging discipline to improve medical information.
Drug discovery is a complex and risky process to find the most effective molecule that have binding affinity to its target and have desired pharmacokinetic profile. It should not have any toxic effects.
1. PHARMACOGENOMICS: Pharmacogenomics can be defined as "the application of genomic approaches and technologies to the identification of drug targets". Using computational tools, we can have an improved understanding of the mechanisms of the diseases and to identify and validate new drug targets. So these highly specific drugs will be more effective as well as will have fewer side effects in comparison to medicines available in the market. With the completion of human genome project, we can search for the genes directly associated with different diseases. It is now possible to understand the molecular basis of incurable diseases and their cure in an efficient manner. Hence, in pharmacogenomics, genetic information is used to predict influence and response of a drug at molecular level. The potential of bioinformatics lies in the identification of useful genes leading to the development of new gene products, drug discovery and drug development.

2. PHARMACOGENETICS. With the development of the field of pharmacogenetics, clinical medicine/ drugs will become personalized. All individuals do not respond equally to the drugs during their treatment; whereas some get cured very early, others have little effect or no response to the drug. Few patients show, hypersensitivity or allergic reactions and some may even show side effects. This is due to the individual's genetic constitution that affects the body's response against the drug. Hence, the application of bioinformatics in the pharmaceutical industry can be very useful to derive personalized medicine. This field is known as pharmacogenetics and can be defined is "the study of how the actions of and reactions to drugs vary with the patient's genes. Much of this variation is known to have a genetic basis". In future. by knowing the, patient's genetic profile doctors will be able to prescribe the best available treatment and drug therapy.

3. GENE THERAPY: Gene therapy approach uses changes in the expression of genes for the treatment of diseases. Presently gene therapy is in its juvenile stage. However, clinical trials for several different Types of cancer and other incurable diseases are under way. In the coming future. The gene therapy approach to treat diseases may become a reality.

4. ANTIBIOTIC RESISTANCE: *Enterococcus faecalis* is a pathogen and cause of bacterial infection among hospital patients. By utilizing bioinformatics, it was revealed that it contains a virulence region made up of a number of antibiotic resistant genes. This region is known as *pathogenicity island,* and could provide useful markers for detecting pathogenic strains. In the future this could be helpful to establish controls prevent the spread of infection. Due to prolonged use of antibiotics, resistance against them had increased in pathogenic bacterial population due to the role of transposon and rapid evolution of R plasmid that may contribute to the bacterium's transformation.

AGROINFORMATICS

The application of bioinformatics in the agricultural science including plant genomes is termed as Agroinfonnatics which concentrates on the sequencing of genomes of plants and animals that can provide benefits for the agricultural community. It can be applied to search for the genes within these genomes to produce stronger, more drought resistant, disease resistant and insect resistant crops. Similarly, it can be helpful to improve the quality of livestock to yield more production by selecting their genes that build them healthier and more disease resistant.
The field of bioinformatics can be an asset for crop improvement. Scientists had completed the sequencing of the genomes of *Arabidopsis thaliana* and *Oryza sativa*. Bioinformatics tools can be used to make comparisons between the numbers, locations and biochemical functions of genes in different plants. By knowing the complete genomes of

important crops and their comparison. it was revealed that the organization of their genes has remained more conserved over evolutionary time. On the basis of the observations, information regarding the model crop systems can be obtained which can be applied for the improvements of other food crops.

(1) By knowing the genomes of the plants, drought resistant varieties, crop varieties capable of tolerating reduced water conditions, soil alkalinity as well as crop varieties capable of tolerating abiotic stress can be generated. That will help in increasing the agriculture land of poorer soil areas, thus helping in more production of grains, cereals, vegetables to combat the increased demand due to population expulsion.

(2) Insect resistant varieties can also be produced using comparative plant genomics. Insect resistant genes can be identified which can increase the ability of the plants to resist insect attack. For example, in *Bt* cotton genes from *Baciltus thuringiensis* have been successfully transferred to cotton that can control a number of serious pests. In this way the enormous amount of insecticide being used in the agriculture can be significantly decreased hence the nutritional quality of the crop is increased and the cost of their production wan lowered.

(3) To improve the quality of livestock yield more production. Several sequencing projects of farm animal, (cow. pl8. Sheep etc.) are undergoing. Completion of these will enhance the understandings of the biology of these organisms. By knowing their genomes and comparison will surely be helpful in improving the production of milk, meat etc. increasing benefits for human nutrition as well as the health of livestock.

PHYLOGENOMICS AND EVOLUTIONARY STUDIES

Phylogeny is the branch of science dealing with the origin und evolution. In past scientists were studying phylogeny and evolution merely on the basis of morphological and anatomical features. But many times it is not possible to achieve concrete results merely on the basis of such traits. After the discovery of karyotyping, scientists have also applied this technology (cytotaxonomy) to understand the phylogeny. Now complete sequence of genomes of many organisms from all three domains of life, eukaryotes. bacteria and archaea arc available. Information on structural, functional and comparative analysis of these genomes and genes from wide variety of organisms will be helpful to understand more about their evolution. The evolutionary studies can be performed to determine the tree of life and the last universal common ancestor. This type of phylogenetic analysis using nucleotide sequences is also known as Molecular taxonomy. PHYLIP is the one of the most popular bioinformatics tool for conducting phylogenetic analysis.

Analyzing and comparing the genetic material of different species is an important method for studying the functions of genes, the mechanisms of inherited diseases and species evolution.

Bioinformatics tools can be used to make comparisons between the numbers, locations and biochemical functions of genes in different organism. By completion of genomes of man and mouse, we have come to know that these are very closely related (>98%) and for the most part we can have comparison between genes in these two species. Hence, mouse can be treated as model for humans. Organisms that are suitable for use in experimental research are termed model organism of the mouse at the molecular level comparisons with man can provide information regarding the gene function, and further this information can be utilized in finding the molecular mechanisms of many human diseases and their treatment.

CREATION OF BIO-WEAPONS

Although bioinformatics has applications, but it is also true that there may be use of harmful aspects interwoven with useful things. After genomic data of pathogenic species is available on the Internet, it can be utilized in the formation bio-weapons. The virus poliomyelitis has been built by scientists using entirely artificial means. US Department of Defense have funded the research as part of a bio-warfare response program. They want to prove the reality of bio-weapons to the world. However, such practice must be discouraged for the benefit of mankind.

LITERATURE RETRIEVAL

The most important application of the bioinformatics is literature retrieval. Scientists researchers are using this facility for designing their problem, what part of the research work has been done documented similar to themselves throughout the globe for comparison of data and publishing. Now a day every press is publishing journals in electronic format which is available online. The literature citation database at the National Center for biotechnology information is called PubMed.

NCBI's numerous database. The PubMed database can be easily searched with Entrez, by a simple keyword search. Full articles are not provided in this database, only citations and abstracts are available to view.

GENOMICS AND PROTEOMICS

Genomics is used to analyze the entire genetic complement of a species. Proteomics is the study of location, structure and function of entire protein content of cell body. During the last decade. bioinformatics has become the powerful tool to explore the treatment of genetic disorders utilizing the genomics approach as well as at the level of proteins, to obtain accurate three-dimensional structures for all known protein families, protein domains or protein folds.

SCOPE OF BIOINFORMATICS

Bioinformatics is the burning buzzword in the field of science and a necessary ally for researchers' in the field of biotechnology. It has become an area of importance for the growth of genomics, proteomics; micro array etc. Without the application of bioinformatics, these areas might have progressed very slowly. It has also become promising field for job seekers. The eventual aspiration of bioinformatics is the innovation of new biological insights that can be applied for the betterment of mankind.

During the last decade there has been an explosive growth in biological data. Due to the improvements in the nucleotide sequence techniques, the contents of nucleotide databases are now getting doubled in size, approximately after every 10 months. The GenBank release has revealed that the nucleotide data has exceeded l74 billion base pairs (2007). This is because the large sequencing projects are producing enormous quantities of nucleotide sequences. To cope with this great quantity of sequenced data, bioinformatics has great promises and relevance to manage and store this data for various forthcoming purposes viz. pharmacogenomics, phylogenomics etc. Bioinformatics has also evolved gradually due to improvement of computational techniques with time. It is not just a useful tool in biological research or drug development, the emergence of bioinformatics as a major thrust area in the field of genomics, proteomics and drug discovery during the last two decades has been well evident. It is advancing rapidly and possessing vast scope in the fields of biotechnology. It is leading the future of drug developers to find cure for genetic disorders and fundamental understanding of living systems particularly after the completion of human genome project. The genetic information which has been revealed after completion of this project can be applied to gene based drug discovery and development. Knowledge of bioinformatics is immensely useful and is now being applied in all the areas of biological science including biotechnology, agriculture, medical sciences, environmental science etc. and is used in the development of:

(1) Molecular medicine for the treatment of incurable diseases.
(2) Discovery of new molecular diagnostic kits.
(3) Producing high yielding, resistant and low maintenance crops.
(4) Environmental benefits in identifying waste cleanup bacteria etc.

Due to researches in various fields of biotechnology, there has been an enormous growth of biological data and services of bioinformatics are being utilized in (l) Storage of this data and (2) Analyzing these huge databases.

Therefore, bioinformatics possesses immense scope and will generate plenty of job avenues for the professionals in this area.