1

HISTORY, AIM AND SCOPE

What is Bioinformatics?

The term *bioinformatics* was coined by Paulien Hogeweg in 1979 for the study of informatic processes in biotic systems.

Bioinformatics is a hybrid branch of biology and information technology or it can be said that it encompasses the knowledge of computer science, statistics, mathematics, chemistry, biochemistry , physics and linguistic. Bioinformatics is the science of developing computer databases and algorithms for the purpose of speeding up and enhancing biological research. The term bioinformatics was first come into use in the 1990s. The branch bioinformatics mainly deals with management, analysis and storage of DNA, RNA and protein sequence data etc. Research in bioinformatics includes methods development for storage, retrieval and analysis of the data. The branch bioinformatics has been variously defined by different scientists and organizations. According to Fredj Tekaia bioinformatics is "the mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information".The National Center for Biotechnology Information (NCBI, 2001) defines bioinformatics as "Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline". The Oxford English Dictionary definition is "Bioinformatics is conceptualizing biology in terms of molecules (in the sense of Physical chemistry) and applying informatics techniques (derived from disciplines such as applied maths, computer science and statistics) to understand and organize the information associated with these molecules, on a large scale" (Luscombe, 2001).The terms bioinformatics, computational biology and bioinformation infrastructure are often used interchangeably, of which bioinformatics focuses on development of practical tools for data management and analysis (e.g. display of genome information and sequence analysis);computational biology encompasses the use of algorithmic tools to facilitate biological analyses and bioinformation infrastructure comprises the collection of information management systems, analysis tools and communication networks supporting biology.

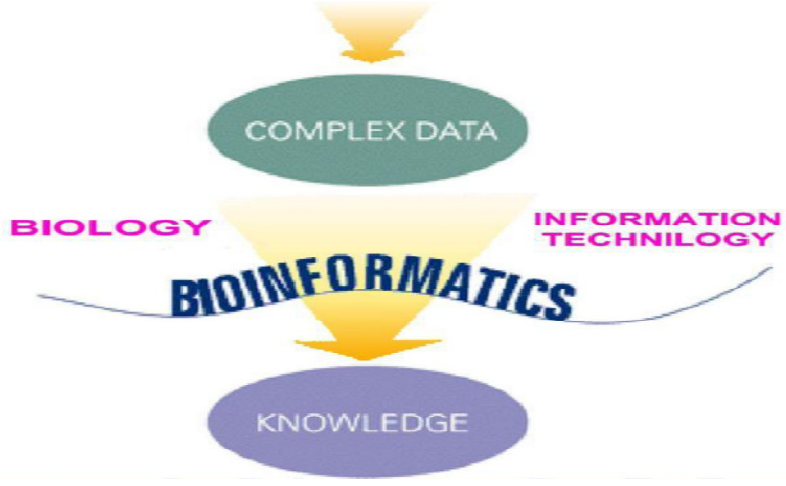

Fig 1. Concepts of Bioinformatics

Historical Background of Bioinformatics

The systematic study of bioinformatics began when Margaret Dayhoff and her collaborators in 1960s at the National Biomedical Research Foundation (NBRF), Washington, D.C., after the development of protein sequencing method by Sanger and Tuppy (1951), organized proteins into families and superfamilies on the basis of degree of sequence similarities. They assembled the databases of protein sequences into a protein sequence atlas and their collection centre became popular as Protein Information Resource (PIR). They also contributed Percent Accepted Mutation (PAM) table for comparison of protein sequences of various organisms. Dayhoff and her associates contributed a lot to modern biological sequence analysis by providing the first protein sequence database and PAM table. Margaret Dayhoff is regarded as founder of bioinformatics. The second landmark in the history of bioinformatics is the creation of DNA sequence databases. In 1974, Theoretical Biology and Biophysics Group founded by George I. Bell at Los Alamos National Laboratory in New Mexico initiated collection of DNA sequence into the GenBank data base to provide theoretical background to experimental work mainly in the field of immunology. The information on protein sequence and DNA sequence stored in data bases were made accessible for public by developing Web pages. An early example of this technology at NCBI was GENINFO developed by D. Benson, D. Lipman and colleagues. Subsequently, a derivative program called Entrez (http://www.ncbi.nim.nih.gov/Entrez) was developed at NCBI. To facilitate accurate data collection the programs Phred and Phrad were developed by Phil Green and colleagues at the University of Washington to assist with reading and processing DNA sequencing data. In 1970, A.J. Gibbs and G.A. McIntyre described a new method – dot matrix method for comparing two amino acids and nucleotide sequences. Although the dot matrix method is useful to determine sequence similarity yet it does not resolve the sequence similarity associated with insertion and deletion. Needleman and Wunsch proposed dynamic programming in 1970 for sequence alignment which could give best alignment of two sequences having match, mismatch, single insertion and deletion. The programme fixed score 1 for every match, 0 for every mismatch and penalty score for every individual gap. These scores were then added across the alignment to get total score for the alignment. The alignment with highest possible score was defined as the optimal alignment. In 1981 Mike Waterman and Temple Smith modified Needleman's and Wuncsh's algorithm and proposed local alignment algorithm. Then Johnson and Doolittle (1986); Thompson and his colleagues (1994) and Notredame and his colleagues (2000) developed multiple sequence alignment program s for aligning three or more sequences at a time. These multiple program alignment helped in tracing evolutionary relationship between organisms and thus initiated evolutionary modeling program. In 1971, Tinoco and his colleagues developed computer based method for predicting RNA secondary structure. Subsequently, Nussinor and Jacobson in 1980, on the basis of algorithm used for aligning DNA and protein sequences developed a fast and efficient computer based method to predict RNA molecules with number of base pairs. This method was further improved by Zuker and Stiegler in 1981. In 1987 database of rRNA molecules was prepared by the laboratory of C. Woese (http://www.cme.msu.edu/RDP.html/inde.html). As time proceeded there was increase in the number of sequenced DNA, RNA and protein. The dynamic programming method developed by Needleman and Wunsch became inefficient to search similarity between huge number of

sequence at a faster rate. To overcome this problem W. Pearson and D. Lipman in 1988 developed a fast computer based program – FASTA. FASTA facilitate similarity search between newly sequenced DNA, RNA and protein with model sequence already present in respective databases. Pearson further improved FASTA program in the year 1990 and 1996. In 1990 a similar program for similarity searching in sequence database called BLAST was developed by S. Attschul and his colleagues. This method is widely used from the web site of the NCBI (http://www.ncbi.nlm.nih.gov/BLAST). BLAST is the most widely used server for searching sequence similarity.Starting from 1970s research had been going on prediction of protein structure. As large numbers of protein structures were determined experimentally computational methods were used to find out proteins shearing similar structural fold. Bowie and his colleagues in 1991 devised a method for searching protein with similar three dimensional conformation. Amos Bairoch (Bairoch*et al*, 1997)developed a method for predicting biochemical activity of unknown protein with known amino acid sequence. With the development of fast and efficient computational method in Feb, 2004 there were more than 22,044 protein structures deposited in the Brookhaven Protein Data Bank (PDB) and 144,731 protein sequence entries in the SwissProt protein sequence database. Complete genome of *Hemophilusinfluenae* was sequenced in the Institute of Genetics Research (TIGR, at http://www.tigr.org/) and was started by Craig Venter. Success of this project had initiated other sequencing projects of various prokaryotic and eukaryotic genomes including Human Genome Project (HGP). As huge numbers of information became available on genome sequence of various organisms, emphasis had been given to generate genome databases. AceDB was a genome database management system developed in 1989 by Jean Thierry-Mieg (CNRS, Montpellier) and Richard Durben (Sanger Institute). Under this system several databases are now found available e.g. TAIR (the *Arabidopsis* Information Resource), SGB (*Saccharomyces* database) etc. that are accessible through the Internet for retrieval of sequence, information about gene and mutants investigators addresses and references. After the sequencing of genome of various organisms annotation of genome was started for determination of number and types of genes, regulatory sites e.g. RNA splicing sites of sequenced genomes. This annotation of genome helped to locate gene in a chromosome. The located gene could be translocated into protein and thus whole genes of a respective genome can be translocated into respective proteins to generate proteome of the organism. These are some of the historical events that have contributed in the development of an exciting branch – bioinformatics.

Starting from Margaret Dayhoff and her collaborators who organized proteins into families and superfamilies on the basis of degree of sequence similarities till today research on different fields of bioinformatics has made tremendous progress in respect to sequencing, similarity search, systematic storage of information in databases and retrieval of stored information, study of phylogeny, drug discovery and design etc. with the discovery of new computerize research tools and technologies. Research contributions by different scientists will give new dimension to the science of bioinformatics in near future.

Some Historical Events on the Field of Bioinformatics

- **1951**- Pauling and Corey propose the structure for the alpha-helix and beta-sheet.

- **1953** - Watson & Crick proposed the double helix model for DNA based x-ray data obtained by Franklin & Wilkins.
- **1954** - Perutz's group developed heavy atom methods to solve the phase problem in protein crystallography.
- **1958** - The first integrated circuit was constructed by Jack Kilby at Texas Instruments.
  1965 - Margaret Dayhoff'sAtlas of Protein Sequences
- **1968** - Packet-switching network protocols were presented to ARPA
- **1970** - The details of the Needleman-Wunsch algorithm for sequence comparison waspublished.
- **1971**- Ray Tomlinson (BBN) invented the email program.
- **1972** - The first recombinant DNA molecule was created by Paul Berg and his group.
- **1973** - The Brookhaven Protein DataBank was announced.
- **1974** - Vint Cerf and Robert Khan developed the concept of connecting networks of computers into an "internet" and develop the Transmission Control Protocol (TCP).
- **1975** - Microsoft Corporation was founded by Bill Gates and Paul Allen.
- **1977** - The full description of the Brookhaven PDB (http://www.pdb.bnl.gov) was published.
- **1978** - The first Usenet connection was established between Duke and the University of North Carolina at Chapel Hill by Tom Truscott, Jim Ellis and Steve Bellovin.
- **1980** - The first complete gene sequence for an organism (FX174) was published. The gene consists of 5,386 base pairs which code nine proteins.
- **1981** - The Smith-Waterman algorithm for sequence alignment was published.
  IBM introduced its Personal Computer to the market.
- **1983** - The Compact Disk (CD) was launched.
- **1986** - The term "Genomics" appeared for the first time to describe the scientific discipline of mapping, sequencing, and analyzing genes. The term was coined by Thomas Roderick.
  The SWISS-PROT database was created by the Department of Medical Biochemistry of the University of Geneva and the European Molecular Biology Laboratory (EMBL).
- **1987**- The use of yeast artifical chromosomes (YAC) was described.
  The physical map of *E. coli*was published.
  Perl (Practical Extraction Report Language) was released by Larry Wall.
- **1988** - National Center for Biotechnology Information (NCBI) created at NIH/NLM EMBnet network for database distribution.
  The FASTA algorith for sequence comparison was published by Pearson and Lupman.
  First annual Cold Spring Harbor Laboratory meeting on human genome mapping and sequencing.
- **1990** - The BLAST program was implemented.
  Molecular applications group is founded in California by Michael Levitt and Chris Lee. Their primary products are Look and SegMod which were used for molecular modeling and protein design.
  InforMaxwas founded in Bethesda, MD. The company's products addressed sequence

analysis, database and data management, searching, publication graphics, clone construction, mapping and primer design.

- **1991** - The research institute in Geneva (CERN) announces the creation of the protocols which made -up the World Wide Web (WWW).
  Myriad Genetics, Inc. was founded in Utah. The company's goal was to lead in the discovery of major common human disease genes and their related pathways.
  Human chromosome mapping data repository, GDB, established
- **1992** -Low-resolution genetic linkage map of entire human genome published.
- **1995** - The *Haemophilusinfluenzea* genomewas sequenced
  1996 - The genome for *Saccharomyces cerevisiae*was sequenced.
  The prosite database was reported by Bairoch, *et.al.*
  Affymetrix produced the first commercial DNA chips.
  Sequence of the human T-cell receptor region was completed.
- **1997** - The genome for *E.coli* (4.7 Mbp) was published.
  NIH NCHGR becomes National Human Genome Research Institute (NHGRI).
  1998 - The genomes for *Caenorhabitiselegans* and baker's yeast were published.
  The Swiss Institute of Bioinformatics was established as a non-profit foundation.
  Inpharmatica, a new Genomics and Bioinformatics company was established by University College London, the Wolfson Institute for Biomedical Research, five leading scientists from major British academic centres and Unibio Limited.
  GeneFormatics, a company dedicated to the analysis and predication of protein structure and function, was formed in San Diego.
- **1999** - First Human Chromosome 22 Sequenced completed.
- **2000** - The genome for *Pseudomonas aeruginosa* (6.3 Mbp) was published.
  The *A.thaliana* genome (100 Mb) was sequenced.
  The *D.melanogaster* genome (180 Mb) was sequenced.
  HGP leaders and President Clinton announced the completion of a "working draft" DNA sequence of the human genome.
  International research consortium publishedchromosome 21 genome, the smallest human chromosome.
  DOE researchers announced completion of chromosomes 5, 16, and 19 draft sequence.
- **2001** - The human genome (3,000 Mbp) was published.
  Human Chromosome 20 sequenced - Chromosome 20 was the third chromosome sequenced completely.
- **2003** -Human Genome Project Completion, April 2003.
  Human Chromosome 14 completely sequenced - Chromosome 14 was the fourth chromosome sequenced completely.
- **2004** - The genome sequence of *Rattusnorvegicus* was completed by the Rat Genome Sequencing project Consortium.
- **2005** – Chimpanzee genome sequenced.
- **2007** – Personal human genome of Dr. C. Venter and Dr. James D. Watson sequenced.

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.

A living cell is a system with cellular components interacting with each other, and these interactions determine the fate of the cell, *e.g.*, whether a stem cell is going to become a liver cell, or a cancer cell. These interacting components include- the genome, the gene transcript and the proteins. Characterization of these three types of components and the associated development of analytical methods lead to the establishment of the three closely related branches of bioinformatics-**Genomics, Transcriptomics and Proteomics.**

Genomics involves extensive analysis of nucleic acids through molecular biological techniques, before the data are ready for processing by computers. Genomics is a science that attempts to describe a living organism in terms of the sequence of its genome (its constituent genetic material).

Proteomics represent the earliest attempt to identify a major subclass of cellular components, the proteins and their interactions. It has been coined from the word “proteome” which is the complete protein complement of a system. Proteomics involves the sequencing of amino acids in a protein, determining its three-dimensional structure and relating it to the function of the protein. Before computer processing comes into the picture, extensive data, particularly through crystallography and NMR, are required for this kind of a study. With such data on known proteins, the structure and its relationship to function of newly discovered proteins can be understood in a very short time. In such areas, bioinformatics has an enormous analytical

and predictive potential. Proteomics also focuses on identifying when and where proteins are expressed in a cell so as to establish their physiological roles in an organism.

Transcriptomics depicts the expression level of genes, often using techniques capable of sampling tens of thousands of different mRNA molecules at a time (eg. DNA microarrays). Transcriptomics has been coined from the word – Transcriptome which is the set of all mRNA molecules (or *transcripts*) in one or a population of biological cells for a given set of environmental circumstances.

Apart from these there are few major branches mentioned below:

Functional Genomics: Since the completion of the human genome, the emphasis has been changing from genes themselves to gene products. Functional genomics assigns functional relevance to genomic information. It is the study of genes, their resulting proteins, and the role played by the proteins.

Cheminformatics: Drug design through bioinformatics is one of the most actively pursued areas of research. Since a great majority of drugs are LMW (Low Molecular Weight) compounds and since many of them are primarily derived from biological sources, there has always been a great interest in the study of LMW compounds of biological origin. Cheminformatics (or chemoinformatics) deals with such compounds, the products of secondary metabolism, often called natural products which has some kind of bioreactivity. This bioactivity can be turned to advantage for therapeutic purposes. Here the expertise of a pharmacologist is required. Cheminformatics involves organisation of chemical data in a logical form to facilitate the process of understanding chemical properties, their relationship to structures and making inferences. Chemical structures are the input to identify similar compounds for screening for biological activity. It also helps to assess the properties of new compounds, by comparison with the known compounds.

Aim of Bioinformatics

Aim of bioinformatics are given here under –

1. Storage of biological data in form of databases to facilitate easy retrieval and submission of new entries of biological information by researchers.
2. Development of tools and resources for data analysis. For example BLASTto find out similar nucleotide/amino acid sequences, ClustalW to align two or more nucleotide/amino acid sequences and Primer3 to design primer probes for PCR techniques etc.
3. Exploitation of computational tools to analyse the biological data and interpret the analysed data in a biologically meaningful manner.

Scope of Bioinformatics

Biotechnology is a new and exciting branch of biotechnology. It mainly deals with software to exploit vast range of biological data stored in databases that are experimentally developed by biotechnologists. Bioinformatics processes data on gene sequencing that is performed by biotechnologists and makes the information accessible for the researchers. This brings biotechnology within the dimension of bioinformatics. Bioinformatics involves molecular biology and computer science. In bioinformatics computer is required to store, retrieve, analyse/predict the composition or structure of the biomolecules. Presently in bioinformatics emphasis has been given on genome research, human genome sequencing, drug design and other disease related issues. It covers important sectors of human life viz. agriculture, health, environment etc.

Scope means area of study. It is obvious that bioinformatics has a very wide scope as given under –

Computational Bioinformatics

It involves computational works to develop an application for dealing biological problems. Computational bioinformatics deals with development of algorithm and software. Algorithm is a logical sequence of steps by which a task can be performed. Molecular biologists, bioinformaticist, computer scientist, system engineer, mathematicians or statistician work together to solve biological problem by designing an algorithm. Another important aspect of computational bioinformatics is database construction and curation. Vast range of information that have generated by biotechnological works on genome, protein and other biomolecules of various organisms are stored in respective databases. The data that are stored in databases are annotated and can be easily retrieved by researchers. There are countless public databases that store different level of biological information.

Application Bioinformatics

Another exciting study area of bioinformatics is application bioinformatics. Application bioinformatics deals with sequence analysis, structureanalysis and function analysis of biomolecules. With the help of computer program bioinformaticist/bioinformatician analyses the sequence of biomolecules like DNA, RNA and protein. Analysis of the sequence of biomolecules is done for evolutionary analysis, identification of mutations, identification of exon and introns, identification and characterization of novel microbes, selection of drug and its appropriate dose for a patient, identification of drug target and gene therapy etc. Structure analysis deals with analysis of structure of unknown protein or nucleic acid by comparing the structure of these unknown biomolecules with the known structures stored in structure databases. It facilitates determination of 2D or 3D structure of proteins and nucleic acids and thus helps in prediction of their function. Function analysis deals with determination of function of gene or protein with the help of function databases. It promotes use of annotated gene or protein for human welfare by biotechnological means.

What is bioinformatics? A proposed definition and overview of the field. NM Luscombe, D Greenbaum, M Gerstein (2001) Methods Inf Med 40: 346-58.