International Journal of Engineering Research and Development
***e**-ISSN: 2278-067X, **p**-ISSN: 2278-800X, www.ijerd.com*
Volume 3, Issue 8 (September 2012), PP. 88-99

Bioinformatics: An introduction and Overview

Mohit Kumar Sharma, Manoj K.Dhar, Sanjana Kaul
Bioinformatics Centre, School of Biotechnology, University of Jammu

Abstract—An extraordinary capital of data is being generated by genome sequencing projects and other experimental efforts to verify and establish the structure and function of biological molecules. The demands and opportunities for interpreting these data are expanding more than ever. Bioinformatics is a science which uses computational techniques to analyze the biological problems; the science of developing and utilizing computer databases and algorithms to accelerate and enhance biological research. Bioinformatics is much more than what this definition says, it's commonly referred as dry lab work which accelerates the wet lab work drastically. Biology + Informatics + Statistics + (Bio-Chemistry + Bio- Physics). *Bioinformatics* ***is a tool to solve the Biological problems based on existing data. It is a method to solve the Biological outcomes based on existing experimental results. It creates the way for the Biologists to store all the data. It makes some lab experiments easy by predicting the outcome of the lab experiment. Sometimes it shows the initial way to start the lab experiment from existing results. It helps the researchers to get an idea about any lab experiments before they start. Computers have become an essential component of modern biology. They help to manage the vast and increasing amount of biological data and continue to play an integral role in the discovery of new biological relationships. This in silica approach to biology has helped to reshape the modern biological sciences. Bioinformatics is a scientific discipline that encompasses all aspects of biological information acquisition, processing, storage, distribution, analysis and interpretation and combines the tools and techniques of biology, physics, chemistry, computer science, information technology and mathematics. Bioinformatics has helped to make possible the current revolution in modern molecular biology. It's method to predict the biological outcomes before anyone go for full fledged research. It's a method to compare the biological data. Ex: sequence analysis. It's a way to predict or solve the protein structure.It the only way for PERSONALIZED MEDICINE in this post genomic era. It's the method to do comparative genomics and predict the Human homolog genes in other species. It's the method to annotate the newly sequenced genomes. We are living in the world of Computers. By analyzing the existing biological data using Information Technology we can predict the biological outcomes. In this review, we provide an introduction to bioinformatics, biological tools, software and databases.***

I. AIMS OF BIOINFORMATICS

The aims of bioinformatics are threefold. First aim of bioinformatics is to organize data in a way that permits researchers to access existing information and to submit new entries as they are produced, e.g. the Protein Data Bank for 3D macromolecular structures [1].While data-curation is an essential task, the information stored in these databases is essentially useless until analysed. Thus the purpose of bioinformatics extends much further. The second aim is to develop tools/software and resources that help in the analysis of data. For example, having sequenced a particular protein, it is of interest to compare it with previously characterized sequences. This needs more than just a simple text-based search and programs such as FASTA [2] and PSI-BLAST [3] must consider what comprises a biologically significant match. Development of such resources dictates expertise in computational theory as well as a thorough understanding of biology. The third aim is to use these tools to analyze the data and interpret the results in a biologically meaningful manner. In bioinformatics, we can now conduct global analyses of all the available data with the aim of uncovering common principles that apply across many systems and highlight novel features.

Data Source	Data Size	Bioinformatics Topics
Raw DNA Sequence	8.2 million sequences (9.5 billion bases)	Separating coding and non-coding regions Identification of introns and exons Gene product prediction Forensic analysis
Protein Sequences	300,000 sequences (~300 amino acids each)	Sequence comparison algorithms Multiple sequence alignments algorithms Identifications of conserved sequence motifs
Macromolecular structure	13,000 structures (~1,000 atomic coordinates each)	Secondary, tertiary structure prediction 3D structural alignment algorithms Protein geometry measurements Surface and volume shape calculations Intermolecular interactions Molecular simulations (force field calculations, molecular movements, docking predictions)
Genomes	40 complete genomes (1.6 million-3 billion bases each)	Characterisation of repeats Structural assignments to genes Phylogenetic analysis Genomic-scale censuses (characterization protein content, metabolic pathways) Linkage analysis relating specific genes to diseases
Gene expression	Largest: ~ 20 time point measurements for ~ 6,000 genes	Correlating expressions patterns Mapping expressions data to sequence, structural and biochemical data
Other data		
Literature	11 million citations	Digital libraries for automated bibliographical searches Knowledge databases of data from literature
Metabolic pathways		Pathway simulations

1. Structural Bioinformatics:

This approach helps in the prediction of 3D structure of protein from its protein sequence. Homology modelling is the best method for predicting the protein structures by using already structured or crystallized protein as a template. MODELLER is one of the best software for Homology modeling [4]. Protein Data Bank is the data base for 3D co-ordinates of a protein. Homology modeling, also known as comparative modeling, is a class of methods in protein structure prediction for constructing an atomic-resolution model of a protein from its amino acid sequence (the "query sequence" or "target"). Almost all homology modeling techniques rely on the identification of one or more known protein structures likely to resemble the structure of the query sequence, and on the production of an alignment that maps residues in the query sequence to residues in the template sequence. The sequence alignment and template structure are then used to produce a structural model of the target. Because protein structures are more conserved than DNA sequences, detectable levels of sequence similarity usually imply significant structural similarity

2. Drug Designing:

It is the process to find the drugs by design based on their biological targets. The field of drug design is a rapidly growing area in which many successes have occurred in recent years. The explosion of genomic, proteomic, and structural information has provided hundreds of new targets and opportunities for future drug lead discovery [5]**.** The process of elucidating the atomic structure of structures and proteins and their complexes and the design of novel, therapeutically relevant ligands based on these structure elucidations, is known as structure based drug design [6]. The majority of drugs are small molecules designed to bind, interact, and modulate the activity of specific biological receptors. Receptors are proteins

that bind and interact with other molecules to perform enormous functions needed for the maintenance of life. They include huge array of cell-surface receptors (hormone receptors, cell-signaling receptors, neurotransmitter receptors, etc.), enzymes, and other functional proteins. Due to genetic abnormalities, physiologic stressors, or some combination thereof, the function of specific receptors and enzymes may become altered to the point that our well-being is diminished.

The drug discovery Schema

As discussed above, the development of any prospective drug begins with years of scientific study to determine the biochemistry behind a medical problem for which pharmaceutical involvement is possible. The result is the determination of specific receptor targets that must be modulated to alter their activity in some way. Once these targets have been identified, the goal is then to discover compounds that will interact with the receptors in some manner. At this initial stage of drug development, it does not matter what effect the compounds have on the targets. We simply wish to find anything that binds to the receptor in any approach.

The modern day drug discovery schema is outlined in Figure 1. The first step is to determine an assay for the receptor. An assay is a chemical or biological test that turns positive when a suitable binding agent interacts with the receptor. Usually, this test is some form of colorimetric assay, in which an indicator turns a specific color when complementary ligands are present. This assay is then used in mass screening, which is a technique whereby hundreds of thousands of compounds can be tested in a matter of days to weeks. A pharmaceutical company will first screen their entire corporate database of known compounds. The reason is that if a successful match is found, the database compound is usually very well characterized. Furthermore, synthetic methods will be known for this compound, and patent protection is often present. This enables the company to rapidly prototype a candidate ligand whose chemistry is well known and within the intellectual property of the company.

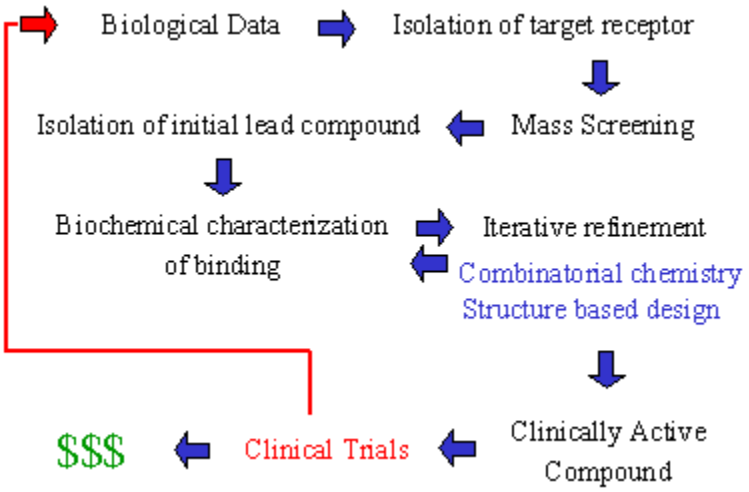


Fig. 1 **Drug discovery Schema.**

Drug design is the approach of finding drugs by design, based on their biological targets. Typically a drug target is a key molecule involved in a particular metabolic or signalling pathway that is specific to a disease condition or pathology, or to the infectivity or survival of a microbial pathogen [5, 6].

Some approaches attempt to stop the functioning of the pathway in the diseased state by causing a key molecule to stop functioning. Drugs may be designed that bind to the active region and inhibit this key molecule. However these drugs would also have to be designed in such a way as not to affect any other important molecules that may be similar in appearance to the key molecules. Sequence homologies are often used to identify such risks.Other approaches may be to enhance the normal pathway by promoting specific molecules in the normal pathways that may have been affected in the diseased state.Computer-assisted drug design uses computational chemistry to discover, enhance, or study drugs and related biologically active molecules.

3. Phylogenetics:

In biology, phylogenetics is the study of evolutionary relatedness among various groups of organisms (e.g., species, populations). The evolutionary history estimated from phylogenetic analysis is usually depicted as branching, treelike diagrams that represents an estimated pedigree of the inherited relationships among molecules, organisms, or both. Sometimes it is called cladistic. Predicting the genetic or evolutionary relation of set of organisms. Mitochondrial SNPs and Microsatellites (DNA repeats) are mostly used in Phylogenetics. MEGA [7],http://paup.csit.fsu.edu/ PAUP [8] are some of the important softwares. Maximum Parsimony and Maximum Likelihood are mostly used methods.). The term phylogenetics is of Greek origin from the terms phyle/phylon, meaning "tribe, race," and genetikos, meaning "relative to birth" from *genesis* Taxonomy, the classification of organisms according to similarity, has been richly informed by phylogenetics but remains methodologically and logically distinct. The fields overlap however in the science of phylogenetic systematics or cladism, where only phylogenetic trees are used to delimit taxa, each representing a group of lineage-connected individuals

Evolution is regarded as a branching process, whereby populations are altered over time and may speciate into separate branches, hybridize together, or terminate by extinction. This may be visualized as a multidimensional character-space that a

population moves through over time. The problem posed by phylogenetics is that genetic data are only available for the present, and fossil records (osteometric data) are sporadic and less reliable. Our knowledge of how evolution operates is used to reconstruct the full tree [9]

4. Computational biology:

Computational biology is an interdisciplinary field that applies the techniques of computer science, applied mathematics, and statistics to address problems inspired by biology. It encompasses the fields of:

- Bioinformatics, which applies algorithms and statistical techniques to the interpretation, classification and understanding of biological datasets. These typically consist of large numbers of DNA, RNA, or protein sequences. Sequence alignment is used to assemble the datasets for analysis. Comparisons of homologous sequences, gene finding, and prediction of gene expression are the most common techniques used on assembled datasets; however, analysis of such datasets have many applications throughout all fields of biology[10].
- Computational biomodeling, a field within biocybernetics concerned with building computational models of biological systems.
- Computational genomics, a field within genomics which studies the genomes of cells and organisms. High-throughput genome sequencing produces lots of data, which requires extensive post-processing (genome assembly) and uses DNA microarray technologies to perform statistical analyses on the genes expressed in individual cell types. This can help find genes of interests for certain diseases or conditions. This field also studies the mathematical foundations of sequencing.
- Molecular modeling, which consists of modelling the behaviour of molecules of biological importance.
- Systems biology, which uses systems theory to model large-scale biological interaction networks (also known as the interactome).
- Protein structure prediction and structural genomics, which attempt to systematically produce accurate structural models for three-dimensional protein structures that have not been determined experimentally.
- Computational biochemistry and biophysics, which make extensive use of structural modeling and simulation methods such as molecular dynamics and Monte Carlo method-inspired Boltzmann sampling methods in an attempt to elucidate the kinetics and thermodynamics of protein functions.

10. Gene Prediction:

Gene finding typically refers to the area of computational biology that is concerned with algorithmically identifying stretches of sequence, usually genomic DNA, that are biologically functional. Predicting the gene by the predefined conditions. Comparative genomics is the best method for predicting the gene.Gene. This especially includes protein-coding genes, but may also include other functional elements such as RNA genes and regulatory regions[11]. Gene finding is one of the first and most important steps in understanding the genome of a species once it has been sequenced.

In its earliest days, "gene finding" was based on painstaking experimentation on living cells and organisms. Statistical analysis of the rates of homologous recombination of several different genes could determine their order on a certain chromosome, and information from many such experiments could be combined to create a genetic map specifying the rough location of known genes relative to each other. Today, with comprehensive genome sequence and powerful computational resources at the disposal of the research community, gene finding has been redefined as a largely computational problem.

Determining that a sequence *is* functional should be distinguished from determining *the function* of the gene or its product. The latter still demands *in vivo* experimentation through gene knockout and other assays, although frontiers of bioinformatics research are making it increasingly possible to predict the function of a gene based on its sequence alone.

Some of the software for Gene Prediction:

GeneMark: GeneMark developed in 1993 was the first gene finding method recognized as an efficient and accurate tool for genome projects [12]. GeneMark was used for annotation of the first completely sequenced bacteria, Haemophilus influenzae, and the first completely sequenced archaea, Methanococcus jannaschii. The GeneMark algorithm uses species specific inhomogeneous Markov chain models of protein-coding DNA sequence as well as homogeneous Markov chain models of non-coding DNA. Parameters of the models are estimated from training sets of sequences of known type. The major step of the algorithm computes a posteriory probability of a sequence fragment to carry on a genetic code in one of six possible frames (including three frames in complementary DNA strand) or to be "non-coding

Genscan: GenScan is an online program to identify complete gene structures in genomic DNA [13]. It is a GHMM-based gene finder for human sequences. GENSCAN was developed by Chris Burge in the research group of Samuel Karlin, Department of Mathematics, Stanford University

12. **Genome Annotation:**

Genome annotation is the process of attaching biological information to sequences. It consists of two main steps:

1. identifying elements on the genome, a process called Gene Finding, and
2. attaching biological information to these elements.

Automatic annotation tools try to perform all this by computer analysis, as opposed to manual annotation which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation pipeline. The basic level of annotation is using BLAST for finding similarities, and then annotating genomes based on that[14]. However, nowadays more and more additional information is added to the annotation platform. The additional information allows manual annotators to deconvolute discrepancies between genes that are given the same annotation.

For example, the SEED database uses genome context information, similarity scores, experimental data, and integrations of other resources to provide the most accurate genome annotations through their Subsystems approach[15]. The Ensembl database relies on both curated data sources as well as a range of different software tools in their automated genome annotation pipeline[16].

Structural annotation consists in the identification of genomic elements.

- ORFs and their localisation
- gene structure
- coding regions
- location of regulatory motifs

Functional annotation consists in attaching biological information to genomic elements.

- biochemical function
- biological function
- involved regulation and interactions
- expression

These steps may involve both biological experiments and in silico analysis. A variety of software tools have been developed to permit scientists to view and share genome annotations. Genome annotation is the next major challenge for the Human Genome Project,now that the genome sequences of human and several model organisms are largely complete[17]. Identifying the locations of genes and other genetic control elements is often described as defining the biological "parts list" for the assembly and normal operation of an organism. Scientists are still at an early stage in the process of delineating this parts list and in understanding how all the parts "fit together" Genome annotation is an active area of investigation and involves a number of different organizations in the life science community which publish the results of their efforts in publicly available biological databases accessible via the web and other electronic means. Here is an alphabetical listing of on-going projects relevant to genome annotation:

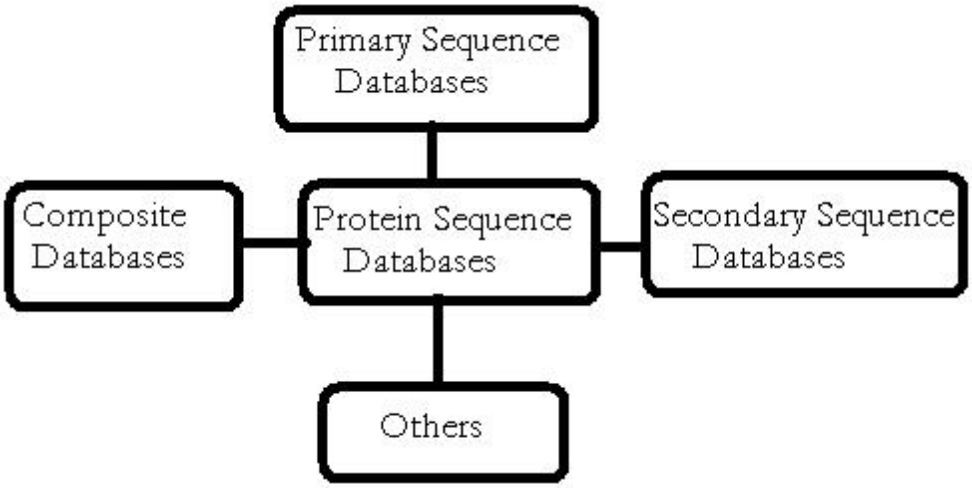
- ENCyclopedia Of DNA Elements (ENCODE)
- Ensembl
- Gene Ontology Consortium
- RefSeq
- Uniprot
- Vertebrate and Genome Annotation Project (Vega)

20. **Database development:**

In some sense Bioinformatics is called as “Comparative Method”. Because Bioinformatics depends on Databases for all of its analysis. So developing data base is a very important project. Many companies surviving by developing and updating the databases. A Computer Database is a structured collection of records or data that is stored in a computer system. The structure is achieved by organizing the data according to a database model. The model in most common use today is the relational model. Other models such as the hierarchical model and the network model use a more explicit representation of relationships[18]. A computer database relies upon software to organize the storage of data. This software is known as a database management system (DBMS). Database management systems are categorized according to the database model that they support. The model tends to determine the query languages that are available to access the database. A great deal of the internal engineering of a DBMS[19], however, is independent of the data model, and is concerned with managing factors such as performance, concurrency, integrity, and recovery from hardware failures. In these areas there are large differences between products.NCBI , PDB and UCSC genome browser are some of the very important databases.
Here's a list of major Biological databases:-

Protein Sequences Databases

1. Primary Protein Sequence Databases:-The primary sequence databases currently hold over 300,000 non-redundant protein sequences. The most commonly-used are SWISS PROT and PIR.
2. Composite Databases:-There are a number of "composite" databases of protein sequences. These compile their sequence data from the primary sequence databases and filter them to retain only the non redundant sequences. The best-known are OWL, NCBI.
3. Secondary Sequence Databases:-Secondary databases are those that contain information derived from the primary sequence databases like PROSITE, PRINTS and Pfam.

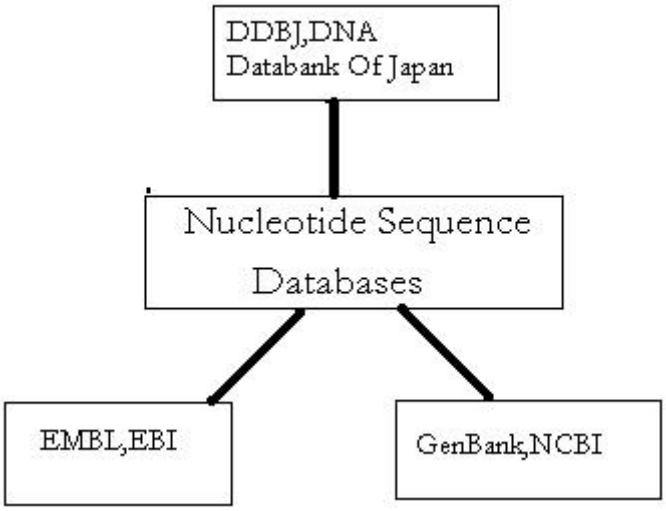


Still there are number of categories in which Protein Sequence Databases can be characterized, these databases are as follows:-

Name	**Type**	**Web Address**
Swiss-Prot	Primary	www.expasy.ch
NCBI Protein database	Composite	http://www.ncbi.nlm.nih.gov/entrez /query.fcgi?db=Protein
PIR-NREF	Primary	http://pir.georgetown.edu/
PROSITE	Pattern based secondary database	http://www.expasy.org/prosite
InterPro	Families/Domains	http://www.ebi.ac.uk/interpro
PRINTS	Family Fingerprints	http://www.bioinf.man.ac.uk/ dbbrowser/PRINTS/
Pfam	Protein families	http://www.sanger.ac.uk/Software/Pfam/
ProDom	Domains	http://www.toulouse.inra.fr/prodom.html
AAindex	Protein property	http://www.genome.ad.jp/aaindex/
PMD,Protein Mutant Database	Literature based information	http://pmd.ddbj.nig.ac.jp/
PRF/SEQDB	Amino acid sequences predicted from genes	http://www4.prf.or.jp/en/
OWL	Composite database	http://umber.sbs.man.ac.uk/dbbrowser/OWL/
SPTR	SWISS PROT+TrEMBL	http://www.hgmp.mrc.ac.uk/Bioinformatics/ Databases/sptr-help.html

Nucleotide Sequence Databases

Name	**Web address**
DNA Data Bank Of Japan,DDBJ	http://www.ddbj.nig.ac.jp
EMBL,EBI Databases	http://www.ebi.ac.uk/embl.html
Genome Databases	http://www.ebi.ac.uk/genomes
UTRdb	http://bighost.area.ba.cnr.it/srs6/
RDP	http://rdp.cme.msu.edu/
rrndb	http://rrndb.cme.msu.edu/
REBASE	http://rebase.neb.com/rebase/rebase.html
DOGS	http://www.cbs.dtu.dk/databases/DOGS/index.html
NCBI	http://www.ncbi.nih.gov/
GenBank	http://www.ncbi.nih.gov/Genbank/GenbankOverview.html
Unigene	http://www.ncbi.nih.gov/UniGene/
Genomes	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db =genome&cmd=search&term=



Structure based Databases

1. Protein Structure Databases

Name	**Web Address**
PDB,Protein Data Bank	http://www.rcsb.org/pdb
PDB-TM	http://www.enzim.hu/PDB_TM/
HOMSTRAD	http://www-cryst.bioc.cam.ac.uk/homstrad
Swiss-Model Repository	http://swissmodel.expasy.org/repository
ModBase	http://alto.compbio.ucsf.edu/modbase-cgi/index.cgi
NRL-3D	http://pir.georgetown.edu/pirwww/dbinfo/nrl3d.html
MMDB	http://www.ncbi.nlm.nih.gov/Structure

2. Nucleic Acid Structure Databases

Name	**Web Address**
NDB	http://ndbserver.rutgers.edu/
RNABase	http://www.rnabase.org/

Databases based on Structure based Classification

Name	**Web Address**
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop
CATH	http://www.biochem.ucl.ac.uk/bsm/cath_new

Organism Specific Databases(Non-Human)

Name	**Web Address**
Rat	http://ratmap.gen.gu.se/
Sheep	http://ws4.niai.affrc.go.jp/dbsearch2/smap/
Mouse	http://www.informatics.jax.org/
Pig	http://www.genome.iastate.edu/
Cow/Cattle	http://ws4.niai.affrc.go.jp/dbsearch2/cmap/ http://sol.marc.usda.gov/genome/cattle/cattle.html
Dog	http://mendel.berkeley.edu/dog.html
Zebra Fish	http://zfish.uoregon.edu/
Horse	http://www.vgl.ucdavis.edu/~lvmillon/
Pufferfish	http://fugu.hgmp.mrc.ac.uk/
Chicken	http://www.genome.iastate.edu/
Mosquito	http://klab.agsci.colostate.edu/
Drosophila	http://flybase.bio.indiana.edu/
E.coli	http://ecocyc.pangeasystems.com/ecocyc/ecocyc.html http://mol.genes.nig.ac.jp/ecoli/

Haemophilus influenzae	http://www.ai.sri.com/ecocyc/hincyc.html
Mycobacterium	http://probe.nalusda.gov:8300/cgi-bin/browse/mycdb
Streptococcus	http://dna1.chem.uoknor.edu/strep.html
Streptomyces	http://www.uea.ac.uk/nrp/jic/gstrgenome.htm
HIV	http://hiv-web.lanl.gov/
Virus Information Resource	http://life.anu.edu.au/./viruses/virus.html

Genome Databases(Human)

Name	Web Address
GDB,Genome Database	http://gdbwww.gdb.org/
GeneCards	http://bioinformatics.weizmann.ac.il/cards/
Gene Map'99	http://www.ncbi.nlm.nih.gov/genemap/
OMIM	http://www.ncbi.nlm.nih.gov/Omim/
TIGR	http://www.tigr.org/tdb/hgi/
GenAtlas	http://bisance.citi2.fr/GENATLAS/

Transcription Related Databases

Name	Web Address
TRANSFAC	http://transfac.gbf.de/TRANSFAC/index.html
TRANSCompel	http://www.gene-regulation.com/pub/databases.html#transcompel
TRED	http://rulai.cshl.edu/tred
JASPAR	http://jaspar.cgb.ki.se
TRRD	http://www.bionet.nsc.ru/trrd/
TESS	http://www.cbil.upenn.edu/tess

21. **Software and Tool development**

Software development is the set of activities that results in software products.Software development may include research, new development, modification, reuse, re-engineering, maintenance, or any other activities that result in software products. Especially the first phase in the software development process may involve many departments, including marketing, engineering, research and development and general management.The term software development may also refer to computer programming, the process of writing and maintaining the source code [20].
Here's a list of major Biological tools and software:

Pairwise Sequence Alignment

- Dot Matrix Plot:-Dot or matrix plots provide an easy and powerful means of sequence analysis, useful for searching out regions of similarity in two sequences and repeats within a single sequence
- ALIGN:-This tool at EBI is very frequently used for both local and global alignments of sequences.
- LALIGN:-Alignments of 2 sequences
- Dotlet:-a Java applet for dotmatrix sequence comparisons
- BLAST 2 Sequences:- This tool produces the alignment of two given sequences using BLAST engine for local alignment.
- SIM:-Alignment tool for protein sequences at Expasy.
- WISE:-Align protein and genomic Sequences at Pasteur.
- JAligner:-Open source Java Implementation of Smith Watermann algorithm.

Multiple Sequence Alignment

- CLUSTAL W at EBI
- CLUSTAL W at embnet
- CLUSTAL W at DDBJ
- GeneBee
- T-Coffee at embnet
- T-Cofee at BioASP
- CLUSTALX:- You can download the program from http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html.
- MultAlin
- MAFFT
- Multiple Sequence Editors:-
  - CINEMA:-Colour Interactive Editor for Multiple Alignment
  - JalView:-Java Multiple alignment editor.

- SeaView
- Strap:-Interactive extendable and scriptable editor for large protein alignments
- GeneDoc

Sequence based homology search tools

- BLAST at NCBI (USA) - basic
- BLAST against (un)finished Genomes at NCBI (USA)
- MEGABLAST at NCBI (large set of DNA query sequences)
- MEGABLAST at Harvard (large set of DNA query sequences)
- PSI BLAST at NCBI (USA) - Position Specific Iterated BLAST
- BLAST at EBI (Hinxton, UK)
- BLAST at GenomeNet (Japan)
- BLAST at BCM: Protein searches
- BLAST at PIR-International Protein Sequence Database
- BLAST search in PRODOM
- OWL BLAST Server
- GeneBee Basic BLAST 2.0
- FASTA at EBI (Hinxton, UK)
- FASTA at NPS (France)
- Fasta 3.3

Protein Sequence Alignment Tools

General Protein Information
Hydrophobicity Plots:-They are useful in predicting membrane-spanning domains, potential antigenic sites and regions that are likely exposed on the protein's surface.
AACompSim is a tool which allows the comparison of the amino acid composition of a Swiss-Prot entry with all other Swiss-Prot entries so as to find the proteins whose amino acid compositions are closest to that of the selected entry
Genetic Code Viewer
PeptideMass Calculate masses of peptides and their post-translational modifications for a UniProtKB/Swiss-Prot or UniProtKB/TrEMBL entry or for a user sequence
CUTTER is tool to generate and analyze proteolytic cleavage.

Motifs/Pattern Searching

NPS@: PATTINPROT search
PROSITE
PRATT
PatScan
PatternFind
Motif Explorer
N-Glycosylation Site Prediction Server
BNL motif searching
MOTIFS in SwissProt at IBCP
PRINTS
BLOCKS
PRODOM
SBASE
MOTIF at GenomeNet (Japan)

Secondary structure Prediction

AGADIR- An algorithm to predict the helical content of peptides
SSCP- Secondary structural content prediction from amino acid composition
GOR
PREDATOR
PredictProtein

Phylogeny Programs

Joe Felsenstein's Phylogeny programs website
Phylogenetic Analysis Computer Programs
Phylogeny software (Glasgow University)
TreeTop - Phylogenetic Tree prediction
CMBI CLUSTAL W
Puzzle: Tree reconstruction for sequences by quartet puzzling and maximum likelihood (Strimmer, von Haeseler)

MacClade Home Page
PAUP
Morkov Chain Monte Carlo - phylogeteic analysis (USA)
Morkov Chain Monte Carlo - phylogeteic analysis (UK)
Morkov Chain Monte Carlo - Molecular clock
TREECON download page (demo version)
Phylogeny server at Pasteur
SOAP Stability of aligned positions
TreeEdit
FORCON download page (sequence format interconversion for the PC only)
Mesquite

Protein Structure Prediction
Comparative Modelling
SWISS-MODEL - An automated knowledge-based protein modelling server
3Djigsaw - Three-dimensional models for proteins based on homologues of known structure
Geno3d - Automatic modelling of protein three-dimensional structure
SDSC1 - Protein Structure Homology Modeling Server
CPHmodels - Automated neural-network based protein modelling server
Threading
3D-PSSM - Protein fold recognition using 1D and 3D sequence profiles coupled with secondary structure information
LOOPP - Sequence to sequence, sequence to structure, and structure to structure alignment
Ab initio
Rosetta Server -Rosetta Server predicts the structure of proteins from the sequence: secondary, local, super secondary, and tertiary.

Molecular Visualization Tools
Rasmol
Swiss PDB Viewer
MolMol
VMD
Cn3d
Pymol

Biological software developed at the Pasteur Institute

- ABCISSE: ABC systems Information on Sequence Structure and Evolution.
- ARIA: Ambiguous Restraints for Iterative Assignment.
- BLAST2TAXONOMY: Blast Taxonomy report.
- CDS: Coding regions.
- CONFMAT: Side chain packing optimization on a given main chain template for protein PDB files.
- COSA: Clustal Output Structural Analysis.
- CYTOSCAPE: Analyzing and Visualizing Networks of Biological Data.
- DECORATE: Side chain packing optimization of a new sequence on a given template main chain.
- ENVIRON: Calculates Energies Associated with Accessible as well as Buried Surface Areas in Proteins.
- EXTRACTCDS: Extract CDS features from a Genbank entry.
- GB2XML: Genbank to XML conversion tool.
- GENEFIZZ: Comparison between genetic and physics segmentations of DNA sequences.
- GENE-LINK: Genetic linkage analysis of experimental backcrosses.
- GMP-TOOL-BOX: GMP-Tool-Box.
- GOLDEN: Fast databanks entries retriever.
- GRUPPI: Clusters of binding sites.
- HOMOLOGY: SCMF Homology Modelling Program.
- HTML4BLAST: Text to HTML blast results formatter.
- ISAPEAKS: Toolbox for data analysis of immune repertoires described by CDR3 usage.
- PISE: A tool to generate Web interfaces for Molecular Biology programs.
- PROSE: Search for Prosite patterns in protein sequences.
- PROTAL2DNA: Align DNA sequences given the corresponding protein alignment.
- RBVOTREE: Report bootstrap values on tree found with Neighbor joining or UPGMA algorithm.
- SEQSBLAST: Extract sequences from a Blast report.
- SIG: Multiple Prosite motifs searching tool.
- SQUIZZ: Sequence/Alignment format checker/converter tool.

- TAXOTRON: Taxotron software: Recognizer, RestrictoScan, RestrictoTyper, Adanson, Dendrograf, AntibioTyper, FactorAna.
- TOPPRED: Topology prediction of membrane proteins (Reimplementation of Gunnar van Heijne algorithm).

II. A FINAL WORD ON BIOINFORMATICS

It is always difficult to present a rapidly moving field such as bioinformatics. Keeping abreast of new developments in bioinformatics is as important an activity as using the data themselves. Current awareness of the field is essential to ensure that all of the relevant available data are captured, maximizing research efficiency. Finally, the best approach to becoming proficient in the use of software tools is often trial and error, and bioinformatics is no exception; trial and error *in silico* can obviate the far less desirable prospect of trial and error in the laboratory, so do not be afraid to experiment with bioinformatics applications—see what the human genome can yield in your hands. Incorporating the usage of Software in Biological analysis is called "Bioinformatics". Bioinformatics is a multidisciplinary field and requires people from different working areas. It is the combination of biology and computer science and is a new emerging field that helps in collecting, linking, and manipulating different types of biological information to discover new biological insight. Before the emergence of bioinformatics, all scientists working in different biological fields, such as human science, ecological science and many other fields, feel a necessity of some tool that helps them to work together. They knew they are all interlinked and had important information for each other, but they did not know how to integrate. In such circumstances, bioinformatics emerges to help these scientists or researchers in fast research and leads to quick inventions by providing readily available information with the help of computer technology. Scientist and researchers spend their whole life in inventing things for human benefits. After so many years of development, they have collected huge amount of valuable data from their experiments all over the world and still this collection is continue and will always continue for the better development of human being. Sometimes, they need to repeat the old research because either it is hard to obtain old data or they do not know whether it exist or not; this wastes their valuable time. Let us take an example of DNA identification. Every species or human beings have particular DNA strands that contain the genetic instructions used in the development and functioning of all known living organisms. By identifying DNA information one can trace generations' links and can find the root of different disease. Earlier it was hard to manage this information. In order to collect and link DNA information from all over the world and to solve many medical complications, bioinformatics is a very helpful hand for them.

In addition, scientists also need a tool that can interlink information from different areas like biology, statistics, genomics etc to make their research faster. For instance, they may need some data regarding effects of particular gene on human being and its effect on animal or on other species, so that they can interlink and generate some beneficial results or antidote that helps in human development. Eventually, bioinformatics provides that help in interlinking information from different fields and leads to quick results.

Finally, bioinformatics also helps in digitizing the information available on paper or in the form of specimen, so that with the help of internet it could be easily available to everyone everywhere. These days, computer is an important part of every research without which so fast development cannot be imagine. Moreover, these days it is important to keep everyone aware of current developments. This will enable everyone to enhance their academic and research skills in their fields at minimum expanse of time, money, and matters. In this scenario, bioinformatics makes information readily available by collecting, linking, and manipulating.

Briefly, bioinformatics is playing a vital role in development of society by providing quick information and making research fast. It is using today's computer technology and biological research together very efficiently. This field is going to generate more opportunities in future for all people working in different areas.

REFERENCES

[1]. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, Rodgers JR, et al. The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur J Biochem 1977;80(2):319-24.,7- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-42

[2]. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85(8):2444-2448

[3]. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-3402

[4]. http://salilab.org/modeller/

[5]. Greer J, Erickson JW, Baldwin JJ, Varney MD (1994). "Application of the three-dimensional structures of protein target molecules in structure-based drug design". *J Med Chem* **37** (8): 1035–1054. doi:10.1021/jm00034a001. PMID 8164249.

[6]. Gubernator K, Böhm HJ (1998). *Structure-Based Ligand Design, Methods and Principles in Medicinal Chemistry*. Weinheim: Wiley-VCH).

[7]. www.megasoftware.net

[8]. paup.csit.fsu.edu

[9]. A.W.F. Edwards & L.L. Cavalli-Sforza (1964). in Systematics Assoc. Publ. No. 6: Phenetic and Phylogenetic Classification: *Reconstruction of evolutionary trees*, 67-76.

[10]. computationalbiology.berkeley.edu/

[11]. Korf I. (2004-05-14). "Gene finding in novel genomes". *BMC Bioinformatics* **5**: 59-67. doi:10.1186/1471-2105-5-59. PMID 15144565

[12]. http://opal.biology.gatech.edu/GeneMark/

[13]. http://genes.mit.edu/GENSCAN.html
[14]. http://blast.ncbi.nlm.nih.gov/Blast.cgi
[15]. http://www.theseed.org/wiki/Main_Page
[16]. http://www.ensembl.org/index.html
[17]. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
[18]. S. Lightstone, T. Teorey, T. Nadeau, Physical Database Design: the database professional's guide to exploiting indexes, views, storage, and more, Morgan Kaufmann Press, 2007. ISBN 0123693896
[19]. Date, C. J. *An Introduction to Database Systems*, Eighth Edition, Addison Wesley, 2000