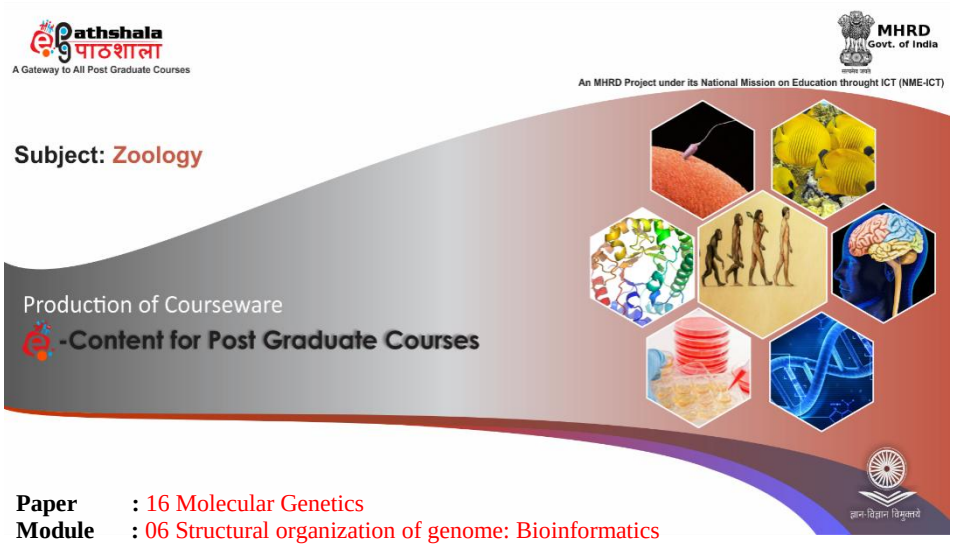


Paper : 16 Molecular Genetics
Module : 06 Structural organization of genome: Bioinformatics

athshala पाठशाला

MHRD Govt. of India

Development Team

Principal Investigator	:	Prof. Neeta Sehgal Department of Zoology, University of Delhi
Co-Principal Investigator	:	Prof. D.K. Singh Department of Zoology, University of Delhi
Paper Coordinator	:	Prof. Namita Agrawal Department of Zoology, University of Delhi
Content Writer	:	Ms. Parul Puri Sri Aurobindo College, University of Delhi
Content Reviewer	:	Dr. Surajit Sarkar, Department of Genetics South Campus, Delhi University

Description of Module	
Subject Name	ZOOLOGY
Paper Name	Molecular Genetics; Zool 016
Module Name/Title	Structural organization of genome
Module Id	M06: Bioinformatics
Keywords	Bioinformatics , in-silico biology, computational biology, systems biology, genomics, pharmacogenomics , epigenomics ,proteomics, metabolomics, high-dimensional biology, CADD -computer aided drug designing, homology, sequence alignment , orthologs, paralogs .

Contents

1. Learning Outcomes

- Emergence of bioinformatics as a discipline.
- Chronological developments in growth and establishment of bioinformatics.
- Importance, Aims and Scope of bioinformatics.
- Understanding key branches and applications of bioinformatics in various fields.
- Role of bioinformatics in rational drug - designing and process of drug development.
- Sequence alignment and homology finding through tools of bioinformatics.

2. Bioinformatics: Brief Historical

Modern day biology is witnessing data explosion with a vast amount of information generated from ongoing genome and sequencing projects. Abundance of data from genome sequences, functional genomics and another high throughput (HTP) techniques with the potential of computing has led to rising of a new discipline namely 'bioinformatics'. In the past few years, bioinformatics has become one of the most developing fields of modern science capable of assimilation, analysis and organization of data from a variety of sources.

Paulien Hogeweg and Ben Hesper coined the term 'Bioinformatics' in 1978 referring to the study of information processes in biological systems. As an interdisciplinary field bioinformatics draws contributions from biology, chemistry, mathematics, statistics and computer science; to understand life and its processes. With the emergence of disciplines such as genetics, biochemistry, molecular biology, and structural biology, the focus of the study of 'life' shifted from the 'macro' properties to 'micro' properties.

Computers have become essential in molecular biology time since protein sequences have become available.The first bioinformatic databases were constructed a few years after the first protein sequence became available. The first protein sequence reported was bovine insulin after the ground breaking work of Frederick Sanger in 1956. Early contributions to bioinformatics embrace comprehensive volumes of antibody sequences released in works of Elvin A. Kabat in 1970. During the journey from the discovery of DNA to be the source of genetic information and elucidation of double-helical arrangement of DNA molecule to the

elucidation of human genome sequence and thereafter, bioinformatics has become an integral part of modern biology.

Foundations of bioinformatics were laid in a breakthrough work by Margaret Oakley Dayhoff appropriately regarded as the 'mother and father of bioinformatics'. A pioneer in the field of bioinformatics' Dayhoff assembled all sequence data information available to create the first bioinformatic database. Dayhoff compiled one of the first protein sequence databases initially published as 'Atlas of Protein Sequence and Structure' in the year 1965. Margaret Oakley Dayhoff pioneered methods of sequence alignment and molecular evolution. Among significant contributions of Dayhoff is the establishment of one-letter code for the amino acids.

The exponential growth in molecular sequence data started in the early 1980s when methods for DNA sequencing became widely available. All of the original databases were organised in a very simple way with data entries being stored in flat files. The data were accumulated in databases such as GenBank, EMBL (European Molecular Biology Laboratory nucleotide sequence database), DDBJ (DNA Data Bank of Japan), PIR (Protein Information Resource) , SWISS-PROT and computational methods were developed for data retrieval and analysis, including algorithms for sequence similarity searches, structural and functional predictions. Such activities of bioinformatics were apparent in the 1980s, although they mainly involved DNA and protein sequence analysis and to a small extent, the analysis of three-dimensional protein structure.

Research in 80s and early 90s focused primarily on development of value-added derived databases to understand the 'sequence - structure - function' relationship.

Late 90s witnessed the transformation of biology into a data-rich science due to development and use of high throughput automated techniques for whole genome sequencing, microarray and proteomics. The use of high-end computational resources and novel algorithms facilitated analysis of these huge datasets leading to our better understanding of bio-complexity. The dawn of new millennium marked the larger integration of bioinformatics with life sciences, a transition accelerating growth of both the disciplines.

The impact of the genome projects of the past 10 years is not simply an increased amount of sequence data, but the diversification of molecular biology data. Automated sequencing has had an enormous impact as it has been at the forefront of the high-throughput generation of various biological data; expressed-sequence tags ESTs and single-nucleotide polymorphisms SNPs among others. Experimental technologies have been developed, notably DNA microarrays for systematically analyzing gene expression profiles and mass spectroscopy for detecting protein-protein interactions.

Development of novel and improved sequence analysis tools accompanied wide-scale sequence data adding to an increase in biological knowledge. Algorithms were simultaneously developed to analyze large datasets, which enabled coding out patterns hidden in the biological data. After the formation of the databases, tools became available to search sequence databases - at first in a very simple way, looking for keyword matches and short sequence words, and then more sophisticated pattern matching and alignment based methods. Since these early efforts, significant advances have been made in automating the collection of sequence information.

Since its introduction rapid but less rigorous sequence database search tool- BLAST has been the mainstay of sequence database searching, complemented by the more rigorous and slower Smith-Waterman and FAST Algorithms. Databases are gathered, organised, disseminated and searched using flat files. With new technologies changes in expression levels of both mRNA and proteins in living cells, both in a disease state or following alteration in external conditions can be directly examined. Patterns of response in cells can be established to provide us to an understanding of the mechanism of action of internal and external factors on a tissue.

2.1. Chronological Developments in Bioinformatics

1902: Emil Hermann Fischer wins Nobel prize for showing that amino acids are linked and form proteins.

1911: Pheobus Aaron Theodore Lerene discovers RNA.

1933: Electrophoresis technique for separating proteins in solution introduced by Tiselius.

1941: George Beadle and Edward Tatum identify that genes make proteins.

1943: first true general- purpose electronic computer (ENIAC) was constructed at the University of Pennsylvania between 1943 and 1946.

1950: Edwin Chargaff finds base pairing rule for cytosine with guanine and adenine with thymine.

1951: First compiler developed by Grace Murray Hopper. Hopper developed the A-0 for the UNIVAC I. She also helped create the COBOL programming language.

Picture 1: Grace Murray Hopper

1951: Linus Pauling and Robert Corey propose α-helix and β-sheet protein structure.

1953: Watson & Crick proposed the double helix structure for DNA based on X-ray crystallographic data obtained by Franklin & Wilkins.

1954: Perutz's group develops heavy atom methods to solve the phase problem in protein crystallography.

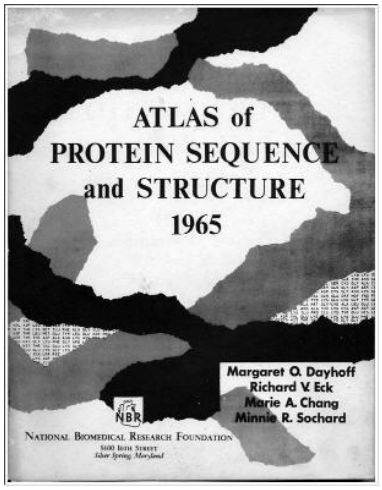
1955: Frederick Sanger analysed sequence of first protein bovine insulin.

1958: First integrated circuit constructed by Jack Kilby at Texas Instruments. Advanced Research Projects Agency (ARPA) formed in US.

1962: Pauling's gave theory of molecular evolution.

1965: Margaret Dayhoff's Atlas of Protein Sequences published.

1966: First bioinformatics system: Margaret Oakley Dayhoff created the first protein sequence database and came up with the PAM model of protein evolution.



Picture 2 & 3: Margaret Dayhoff's - Atlas of Protein Sequences (1965)

1968: Packet-switching network protocols are presented to ARPA.

1970: details of Needleman-Wunsch algorithm for sequence comparison published.

1971: E-mail program invented by Ray Tomlinson.

1972: first recombinant DNA molecule created by Paul Berg, Herbert Boyer, and Stanley N. Cohen.

1973: Brookhaven Protein DataBank announced. Robert Metcalfe from Harvard University describes '*Ethernet'* in his Doctoral thesis.

1974: Vinton Gray 'Vinton' Cerf and Robert Elliot Kahn developed the concept of connecting networks of computers into an 'internet' and develop Transmission Control Protocol/Internet protocol; TCP/IP. Specification of Internet Transmission Control Program by Vinton Cerf, Yogen Dalal and Carl Sunshine, Network Working Group contains first use of the term *internet*, as shorthand for *internetworking.*

1975: Microsoft Corporation is founded by Bill Gates and Paul Allen. Two-dimensional electrophoresis for separation of proteins on SDS -PAGE is combined with separation according to isoelectric points by P. H. O'Farrell.

1976: Unix-to-Unix Copy Protocol developed at Bell Labs. E. M. Southern published details of Southern Blot technique of specific sequences of DNA.

1977: Allan Maxam and Walter Gilbert; Frederick Sanger reports methods for DNA sequencing.

1980: complete gene sequence of first organism, a single stranded bacteriophage φX174 published. Multi-dimensional NMR for protein structure determination described by Wuthrich et. al. Intelli Genetics Suite of programs for DNA and protein sequence analysis developed.

1981: Smith-Waterman algorithm for sequence alignment is published. IBM introduces its Personal Computer.

1982: Genetics Computer Group (GCG), created as a part of the University of Wisconsin, of Wisconsin Biotechnology Center. Gen Bank Released.

1983: Production of DNA clone (cosmid) libraries by Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL).

1984: Jon Postel's Domain Name System placed on-line. Macintosh announced by Apple Computer.

1985: FASTP / FASTN algorithm published.

'Genomics' coined by Thomas Roderick appears for the first time to describe the scientific discipline of mapping, sequencing, and analysing genes.

SWISS-PROT database created by Department of Medical Biochemistry, University of Geneva and European Molecular Biology Laboratory EMBL.

PCR reaction is described by Kary Mullis and co-workers.

1986: automated sequencing technique by Leroy Hood.

1987: Use of YAC's yeast artificial chromosomes described by David T. Burke and coworkers.

Physical map of *E. coli* is published by Y. Kohara and coworkers.

PERL - Practical Extraction Report Language released by Larry Wall.

1988: National Centre for Biotechnology Information, NCBI created at NIH / NLM EMB net network for database distribution.

FASTA algorithm for sequence comparison is published by Pearson and Lipman. Telomere sequence having implications for aging and cancer research is identified at LANL. Human Genome Initiative is started.

1990: BLAST program is implemented. InforMax is founded with company's products address sequence analysis, database and data management, searching, publication graphics, clone construction, mapping and primer design.

1991: CERN research institute in Geneva announces the creation of the protocols which constitute the World Wide Web. Linus Torvalds announces a Unix-Like operating system which later becomes Linux creation. Use of expressed sequence tags ESTs described.

Human chromosome mapping data repository, Genome Database GDB is established.

1992: Low-resolution genetic linkage map of entire human genome published. Guidelines for data release and resource sharing announced by DOE and NIH.

1993: International IMAGE Consortium established to coordinate efficient mapping and sequencing of gene-representing cDNAs.

1994: Netscape Communications Corporation founded; releases a commercial version of NCSA's Mozilla.

PRINTS database of protein motifs is published by Attwood and Beck.

EMBL-EBI European Bioinformatics Institute established, Hinxton, UK.

Completion of second-generation DNA clone libraries representing each human chromosome by LLNL and LBNL.

1995: Microsoft releases version 1.0 of Internet Explorer. Sun releases version 1.0 of Java. Sun and Netscape release version 1.0 of JavaScript.

First non-viral whole genome sequenced for the bacterium *Haemophilus influenzae.*

Sequence of smallest bacterium, *Mycoplasma genitalium*, completed; provides a model of the minimum number of genes needed for independent existence.

Physical map with over 15,000 STS markers published.

1996: *Saccharomyces cerevisiae* genome sequence completed.

PROSITE database is reported by Bairoch et.al.

Affymetrix produces the first commercial DNA chips.

The sequence of the human T-cell receptor region completed.

Archaebacteria- *Methanococcus jannaschii* genome sequenced; confirms the existence of third major branch of life on earth.

1997: genome for *E. coli* published.

1998: genomes of *Caenorhabitis elegans* and baker's yeast are published.

Swiss Institute of Bioinformatics is established as a non-profit foundation.

Craig Venter forms Celera Genomics in Rockville, Maryland.

1999: First Human chromosome 22 completely sequenced.

2000: *Pseudomonas aeruginosa* genome published.

Arabidopsis thaliana genome sequenced.

Drosophila melanogaster genome sequenced.

International research consortium publishes chromosome 21 genome, the smallest human chromosome and the second to be completely sequenced.

2001: Human genome published. Human Chromosome 20 completely sequenced.

2002: genome sequence of common house mouse 2.5 Gb published.

2003: Human Genome Project completed.

2004: *Rattus norvegicus* Brown Norway laboratory rat draft genome sequence completed.

2.2. Defining 'Bioinformatics'

Bioinformatics is an application of computer technology to manage biological information. Bioinformatics is realizing biology in terms of macromolecules and then applying 'informatics' techniques (derived from disciplines such as computer science, statistics and applied mathematics) to understand and organize the information associated with these molecules on a large-scale.

According to Oxford English Dictionary definition 'Bioinformatics' is the science of information and information flow in biological systems, especially of the use of computational methods in genetics and genomics. Bioinformatics thus is the science of storing, extracting, organising, analysing, interpreting and utilising information from biological sequences and molecules.

The primary goal of bioinformatics is to increase the understanding of biological processes**.** It has been mainly fuelled by advances in DNA sequencing and mapping techniques. Over the past few decades rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. A major activity in bioinformatics is to develop software tools to generate useful biological knowledge. Bioinformatics is similar to computational biology and has similar aims to it but differs at the level that it works with basic biological data (e.g. DNA bases) and uses computers to better understand details of biological processes, computational biology is a subfield of computer science which builds large-scale general theoretical models of biological systems seeking to expand our understanding of them from an abstract point of view. Bioinformatics relies substantially on significant contributions made by scientists in various fields, including but not limited to, biology, chemistry, mathematics, computer science and statistics. Bioinformatics has evolved into a full-fledged multidisciplinary subject that integrates developments in information and computer technology as applied to biotechnology and biological sciences. Bioinformatics uses computer software tools for database creation, data management, data warehousing, data mining and global communication networking. Bioinformatics enables recording, annotation, storage, analysis, and searching/retrieval of nucleic acid sequence (genes and RNAs), protein sequence and structural information. This includes databases of the sequences and structural information as well methods to access, search, visualize and retrieve the information.

2.3. Computerised biological information

2.3.1. Sequential Flow of information in biology

Francis Crick in 1958 described flow of information in the mechanism of protein synthesis. DNA, RNA and proteins, are sequential linear biopolymers. The sequence of their monomers effectively encodes information. According to Francis Crick's hypotheses about the relation of genes to proteins, the linear order of bases in DNA determined the corresponding linear order of amino acids in proteins.

The Central Dogma deals with the detailed base to base transfer of sequential information from nucleic acids (DNA, RNA) to proteins.

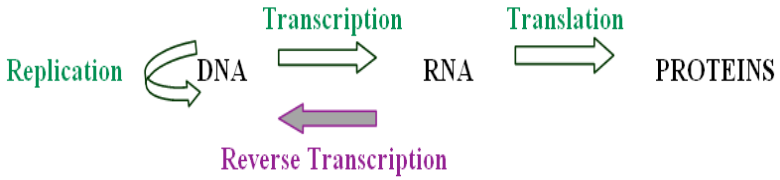


Fig. 1: Information flow in biological systems

DNA can be copied to DNA through ***replication*** by DNA polymerase;

DNA information can be copied into mRNA via ***transcription*** by RNA polymerase and transcription factors. Mature mRNA finds its way to a ribosome, where it is translated.

Proteins can be synthesized using the information in mRNA as a template by process of ***translation***.

Hence, biological information flow in terms of mechanisms operating with the transcription of DNA to mRNA through an enzyme that opens the weak H-H hydrogen bonds of the double helix of DNA and facilitates the hydrogen bonding of complementary bases of RNA along a linear stretch of DNA. The order of the bases in DNA is the information that determines the sequence of the bases in RNA. The messenger RNA then moves to the ribosomes where the translation machinery decodes the information to order the amino acids of the protein.

➢ **Special transfers** are known to occur only under specific conditions in underlying cases.

Reverse transcription the transfer of information from *RNA to DNA* (the reverse of normal transcription) known to occur in the case of retroviruses, such as HIV, as well as in eukaryotes, in the case of retro-transposons and telomere synthesis.

RNA replication the copying of *one RNA to another*. Many viruses replicate this way. The enzymes that copy RNA to new RNA, called RNA-dependent RNA polymerases, are also found in many eukaryotes where they are involved in RNA silencing.

Direct translation from *DNA to protein* has been demonstrated in a cell-free system, using bacterial cell extracts containing ribosomes. These cell fragments could express proteins from foreign DNA templates, and neomycin was found to enhance this effect.

Methylation-Variable methylation states of DNA alter gene expression levels significantly. Variation in methylation occurs through the action of enzyme DNA methylases. The effective information content has been changed by means of the actions of a protein or proteins on DNA, but the primary DNA sequence is not altered. If heritable the change is considered to be 'epigenetic'.

Prions are proteins that propagate themselves by making conformational changes in other molecules of the same type of protein. This is not an exception to the central dogma since the protein sequence remains unchanged

2.3.2. Computer Fundamentals

2.3.2.1. Programming languages in bioinformatics

Essentially, a source of biological data and an appropriate programming language are indispensable elements to provide a bioinformatics-based solution to a biological problem.

A **programming language** is a standard language for communicating instructions to the computer. Programming languages are designed to organize and express complex algorithms that enhance computing. Most popular use of computer languages in biology finds a way through their application in bioinformatics.

Programming languages require a greater level of specificity. Nearly every programming language has the potential to be used in bioinformatics. Popular languages relevant to bioinformatics are Perl, Python, Java, C, C++ and C#. These are used in developing widespread bioinformatics applications. Other languages such as Ruby, PHP, R ,SQL also find their usage in bioinformatics.

Perl stands for **P**ractical **E**xtraction and **R**eporting **L**anguage. Perl is the most established language in bioinformatics and is available as a collection of Perl modules - BioPerl used for

bioinformatics applications. Ensembl an automated genome-annotation project at European Bioinformatics Institute EBI has coding based on BioPerl.

C and C++ - C is the ideal programming language for operating systems. C is one of the oldest programming languages still in popular use today. At NCBI currently-maintained implementation of BLAST is part of the NCBI C++ toolkit.

2.3.2.2. Role of Supercomputers in Biology

The unprecedented boom of biological data generated from various sequencing experiments has witnessed a turning point for biology's use of computers. Computers are classified on the basis of increasing computing power and efficiency as Microcomputers (Personal computer, laptop, netbook, tablet computer, smart phones) > Minicomputers > Mainframe computers > Supercomputers.

Microcomputers are single user computers or personal computers, whereas minicomputers and mainframe computers are multi-user based. Mainframe computers are capable of supporting hundreds to thousands, of users simultaneously. A supercomputer can execute a single application program much faster than a mainframe. A supercomputer is a computer with an enormous speed of calculation and memory. The processing speed of supercomputers are measured in floating point operations per second; FLOPS.

Ongoing state-of-art biological research imposes a constant challenge to integrate, search, analyze, organize, compare and share vast volumes of diverse biological information. Supercomputers with their ability to store, access and analyze large amount of data have become dominant choice to accomplish high-performance computing in biology. Supercomputers open up new horizons, offering the possibility of discovering new ways to understand biocomplexity.

Supercomputing has enabled bioinformaticians to explore through the sequences looking for patterns and similarities. Expanding proficiency in software development, simulation and modeling of cellular components, parallel computing, mass storage of biological database and data integration that provide a powerful blend to speed up discovery for bioinformatics and

computational biology research. Supercomputers have also made possible entirely new fields of study, such as whole-genome comparisons, protein folding and protein-protein interactions inside the cell. Large scale computing environment can effectively simulate the behaviour of individual proteins. Detailed sub-cellular simulations such as manufacture of proteins on ribosome allow researchers to mimic experiments *in silico*. Researchers use supercomputers to study orderly folding of proteins, their binding properties, protein-ligand interaction (including drug) to search possible drug targets that can facilitate effective drug designing.

2.4. Branches of bioinformatics

Bioinformatics deals at the level of all-inclusive understanding and expression of macromolecules of life. It deals with holistic view of the molecules that make up a cell, tissue or organism aiming primarily at detection of genes, mRNA, proteins and metabolites.

Collective study of biological entities (gene, RNA and proteins) is based on-omics technologies. Genomics, transcriptomics and proteomics are believed to be three cardinal branches of bioinformatics.

Genome is complete set of genetic information of an organism. Genetic information is stored in DNA in form of genes. The study of the structure, function and expression of all the genes comprising an organism is called '**genomics**'.

Gene expression patterns and gene function can also be influenced by reversible, non-genetic changes in genome structure or chromosome organisation without altering the DNA sequence. Such 'epigenetic' changes involve DNA methylation and modification of histones by methylation, acetylation and reversible phosphorylation can be studied through **epigenomics**.

Common variations in DNA sequences such as single nucleotide polymorphisms (SNPs), may have functional significance if the changed codon codes for different amino acid. SNP profiling has a significant role in pharmacogenomics. **Pharmacogenomics** describes the influence of an individual patients DNA-sequence variation on the effectiveness of a drug.

Sum total of RNA transcribed in a cell or organism under given set of conditions is its **transcriptome.** It is the language of expression of genes on DNA. **Transcriptomics** enables characterisation of the entire mRNA content along with intended changes in gene expression profile under altered physiological state of the cell, tissue. Expressed sequences of DNA can be used as tags ESTs, that can be used as probes to determine the presence or absence of similar transcripts in other tissues. Vast amount of EST data is available in databases such as GenBank.

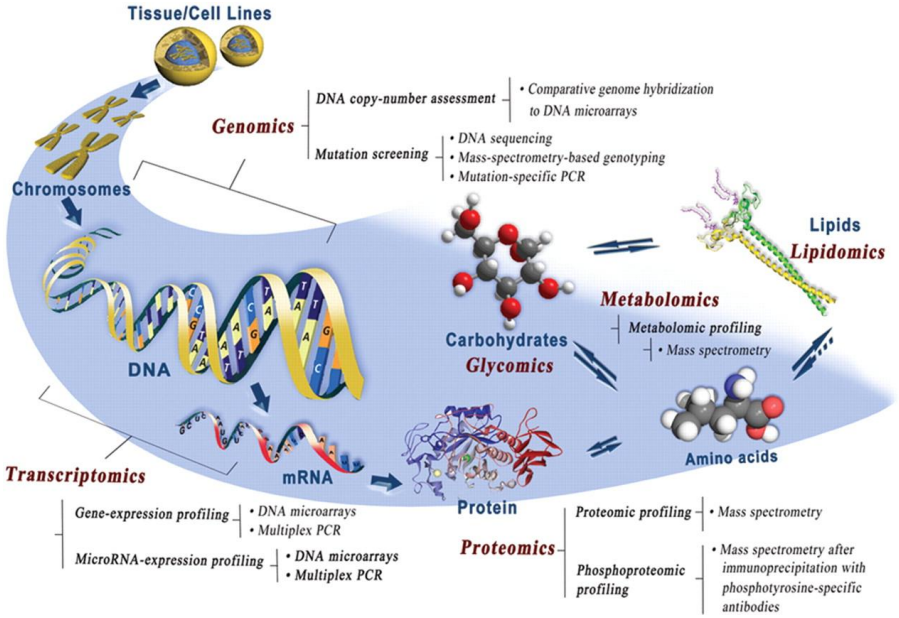


Fig.2: Cardinal Branches of bioinformatics and their applications.
Reproduced fromWu et al.(2011), Journal of Dental Research ,90 : 5; 561-572.

The localisation, functions, modifications and amount of entire proteins produced at organism level is ascertained through **proteomics**. Set of protein-to-protein interactions in a cell - that is its **'interactome'**; is dynamic. Many interactions are transient, and variable in terms of phase of cell development and cellular localisation. Compared to genome; interactome provides deeper insight into resulting biological complexity and allows comprehensive understanding of biosystems. Additionally, knowledge of collective metabolic profile of a living system under specified conditions (= **metabolome)** may help predict metabolic response of organism to changing physiological conditions. A measure of the metabolic

fingerprint of physio-chemical perturbations caused in state of disease, drugs toxicity is 'metabo**n**omics'.

Simultaneous examination of the genomic variations, mRNA transcripts, proteins, and metabolic profile of an organ, tissue, or an organism in various physiological states is referred to as **High-dimensional biology (HDB).** Collaborative study of biological system based on modelling and discovery of emergent properties of cells, tissues and organisms functioning as a system has given way to new discipline **systems biology**, a subset of bioinformatics.

2.5. Importance of bioinformatics

Bioinformatics is concerned about the creation and maintenance of databases of biological information whereby researchers can both access existing information and submit new entries. Biomolecular structure, functional genomics, proteome analysis, cell metabolism, biodiversity, drug and vaccine design are some of the areas in which bioinformatics is an integral component.

Analyses in bioinformatics predominantly focus on three types of large datasets available in molecular biology: macromolecular structures, genome sequences, and the outcomes of functional genomics experiments (e.g. expression data). Additional information includes the submitted scientific findings and "relational data" from metabolic pathways, phylogenetic analysis, molecular simulations and protein-protein interaction networks. Bioinformatics employs a wide range of computational techniques including sequence and structural alignment, database design and data mining, macromolecular geometry, phylogenetic tree construction, prediction of protein structure and function, gene finding and gene annotation. The emphasis is on approaches integrating a variety of computational methods and heterogeneous data sources. Bioinformatics finds some representative applications such as homologue finding and drug designing. Functional genomics, proteomics, discovery of new drugs and vaccines, molecular diagnostics and pharmacogenomics are some of the areas in which bioinformatics has become an integral part of Research & Development.

2.5.1. Biological Databases

Biological database is an assemblage of data which is structured, searchable, timely updated, validated and cross- referenced. Main purpose of databases is to make biological data available and systemised; for easy retrieval as well as analysis of computed biological information.

Biological databases have following seven attributes:

- **Data heterogeneity** refers to diversity in type of data obtained as sequences, three dimensional structures, patterns, graphic representations et cetera.
- **High volume data** indicates large amount of information hidden to be discovered from the data.
- **Uncertainty** as associated with any biological phenomena to be true.
- **Data Curation** involves maintenance, management and value addition to the obtained biological data.
- **Large scale data integration** is needed due to newly generated data and additional information discovered on existing data through ongoing researches.
- **Data sharing** exchange of data through databases among scientific community.
- **Dynamic and subject to change** with scope of addition and improvement of data.

2.5.1.1. Classification of biological databases

Biological databases are broadly classified on following basis:

a. Based on Data type:

1. Genome database:

- **Ensembl** - genome databases for vertebrates and other eukaryotic species
- **Xenbase** - *Xenopus* web resource
- **ZFIN** - Zebrafish Information Network
- **WormBase** - biology and genome of *C. elegans*

2. Sequence database:

- **Nucleotide database -** EMBL (European Molecular Biology Laboratory), Genomes server, GenBank, DDBJ. (DNA Data Bank of Japan)
- **Protein database:** Swiss-Prot, TrEMBL, InterPro, PANDIT.

3. Structure database : PDB, DALI, NDB.

4. Microarray database : MIAME, ClustArray.

5. Chemical database : ChEMBL, ChEBI.

6. Pathway database : KEGG, BioCyc.

7. Enzyme database : IntEnz, REBASE, ExPASy .

8. Disease database : OMIM (Online Mendelian Inheritance in Man) and OMIA.

9. Literature database: MEDLINE.

b. Based on Maintenance status : NCBI, EMBL, SIB.

c. Based on Data access: Publicly available ,available with copyright, browsing only; accessible but not downloadable, academic but not freely available ,proprietary commercial, restricted SQL queries under database management system.

d. Data source :

- **Primary database** also called archival. Contains original data from the researchers .Public or open access. Identification of sequences of interest from primary databases involves screening a large number of entries. Examples of primary databases includes GenBank, EMBL and DDBJ for DNA and RNA sequences, SWISS-PROT and PIR (Protein Information Resource) for protein sequences and PDB (Protein Data Bank) for molecular structures.

- **Secondary database** is curated database that contains additional information derived from analysis of entries from primary database. SCOP (Structural Classification of Proteins) describes structural and evolutionary relationships between proteins of known structures; CATH (Class, Architecture, Topology, Homology) which includes a hierarchical classification of protein structures are examples of secondary databases.

- **Composite Databases** represent an amalgamation of several primary database sources. Composite databases allows a user to access all the relevant information from a single source without the need to search every primary database .NCBI (National Centre for Biotechnology Information) is an example of widely used composite database, which includes several primary and secondary databases such as GenBank, PubMed, OMIM.

- **Integrated Databases** contain data that has been collected from different, but related organisms. Such data are very useful for comparative genomics studies and provide a better insight into the evolutionary relationships between the genomes of different organisms. ATIDB (*Arabidopsis thaliana* Integrated Database) provides a comparative data of genome and transcriptome sequences between the model organism, *Arabidopsis thaliana* and related Brassica species such as *B. rapa, B. nigra*.

e. Database design: relational and object-oriented.

f. Organism: Human, bacteria, virus, plant et cetera. Few organism-specific databases are listed as:

Maize GDB	:	corn/maize (*Zea mays)*
FLYBASE	:	*Drosophila*
Oryzabase	:	rice species (*Oryza* species)
TAIR	:	The Arabidopsis Information Resource
OMIM	:	Human genes and genetic disorders

3. Scope and Applications of bioinformatics

Bioinformatics databases and tools have changed the pace of basic and applied research. Applications of bioinformatics are widespread among major thrust areas of genomics, molecular biology, biotechnology, biomedical sciences, agriculture, environmental and pharmaceutical sciences. These include numerous applications in the field of gene therapy, molecular medicine, preventive medicine, drug development, microbial genome applications, forensic analysis, evolutionary studies, crop improvement (to improve nutritional quality of plants and development of drought resistant varieties), development of alternative sources of

energy, waste clean-up, bio-weapon creation, phylogenetic analysis, comparative studies and computational biology.

At its larger prospective, security of data generated and transferred, validity of data and data accuracy are also imperative research themes of bioinformatics.

3.1. Computer aided drug design - CADD

Remedial treatment of diseases requires development of therapeutically important molecules that can serve as potential drug compounds. Drug discovery and development methods thus play crucial role in designing pharmaceutically important compounds which can function as therapeutic agents i.e. drugs.

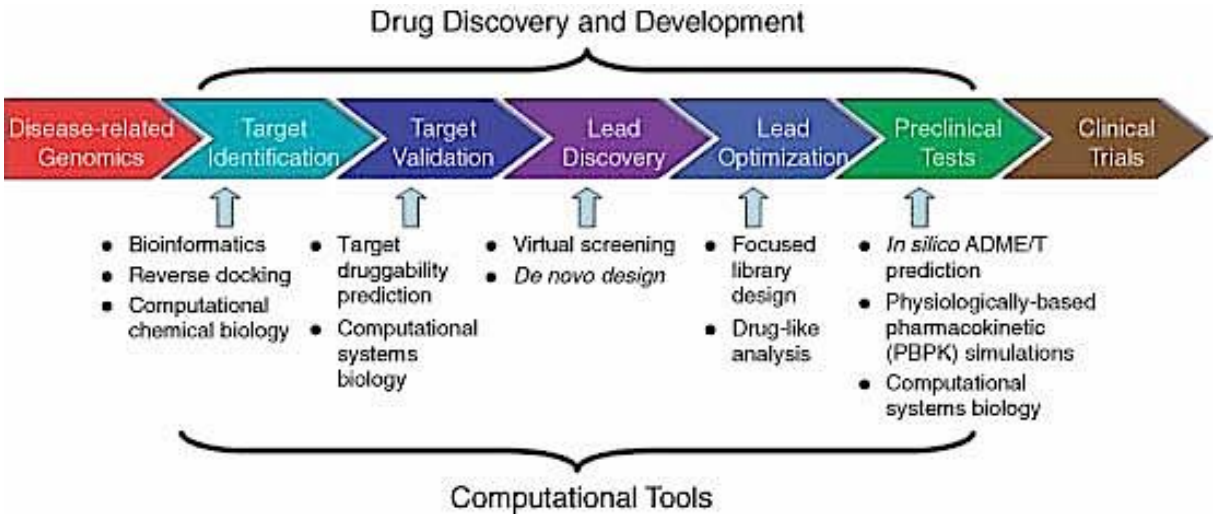


Fig. 3: Computational tools in drug discovery and development process
(Li et al., 2008, Wiley Encyclopedia of Chemical Biology).

Computational techniques have paved way for implementation of bioinformatics principles in drug discovery and designing process. The research area of drug designing using computer based approach is designated as **rational drug design, computer-aided drug design (CADD).**

Strategies for CADD vary depending on the extent of structural and other information available regarding the target (enzyme / receptor) and the ligands. Computer aided drug design, CADD methods are broadly categorised into ***structure-based drug design and***

ligand-based drug design approaches. Structure (target)-based approach or direct drug designing requires structural information of the target which can be obtained from NMR, X-ray crystallography, or homology modelling. Structure based approaches include ligand docking, pharmacophore and ligand design methods.

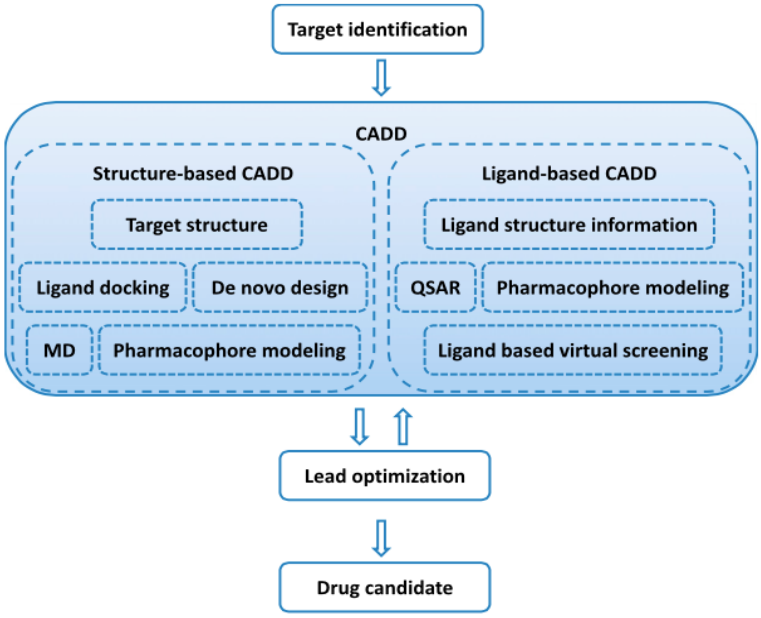


Fig. 4: Strategical approaches in drug designing -'direct' and 'indirect' method.
Sliwoski et al., Pharamacological Reviews, 66 (1):334-395.

Ligand-based methods are indirect methods that rely on ligand structure information to deduce target properties. The proficiency of CADD in the process of drug discovery is to accelerate target, hit identification (active drug candidates); finding lead (most likely candidates for further evaluation) compounds and lead optimisation (transforming bioactive compounds into suitable drugs by improving their physicochemical and pharmaceutical properties) for following preclinical and clinical trials.

3.2. Systems biology

Understanding complexity of a system, how it functions (under variable conditions) and interacts with its counterparts can provide deeper insight about life phenomena. Systems biology as a subdiscipline of bioinformatics allows *in silico* reconstruction of the pathways

and cellular networks to exemplifying functioning of system components in association. It aims at understanding molecular systems and networks' functioning in integrated manner. A significant task of systems biology is to build models of biosystems to better understand system dynamics and behavior. Models accounting for circadian clocks, cell cycle, models simulating transcription and translation and larger network simulation such as microbial networks to understand their interaction with environment, have been developed as an application of systems biology. Moreover various cellular models have allowed to better understand mechanisms underlying many genetic and neurological disorders.

3.3. Application of bioinformatics in genome organisation: homology-homologous, orthologous & paralogous sequences

Genome organisation refers to the sequential organisation of the simple and complex genomes. Genome of organisms contains all necessary information ciphered-in an arranged sequence of universal nucleotide bases A, T (U), G, C. An organism's genome gives a complete set of specification of the organism and; its activity at the molecular level at any moment depends primarily on the amounts and distribution of its gene products. Understanding of organisation of genomes help understand distribution, arrangement, expression and regulation of genes-gene products ,regions contributing to gene expression - coding regions (exon),distribution of non -coding intronic regions (if present; as in eukaryotes), regions conserved among and within species, presence of gene duplications, mutations, pseudogenes; providing an overall picture of how and why is the organism as though. The basic mechanisms for evolution are mutation, recombination and natural selection. These processes are closely related to genes and dynamics of their replication and translation. Identifying genes and dynamics of genes across species is generally accomplished through homology searches.

Homology generally means relationship of nucleic acid or protein sequences that are descent from a common ancestral sequence. The concept of homology – common evolutionary ancestry is central to computational analysis. Homology can be inferred from results of sequence similarity. Two sequences are homologous if they share a common evolutionary

ancestry. **Similarity** searching is effective and reliable because sequences that share significant similarity can be inferred to be homologous that is; they share a common ancestor.

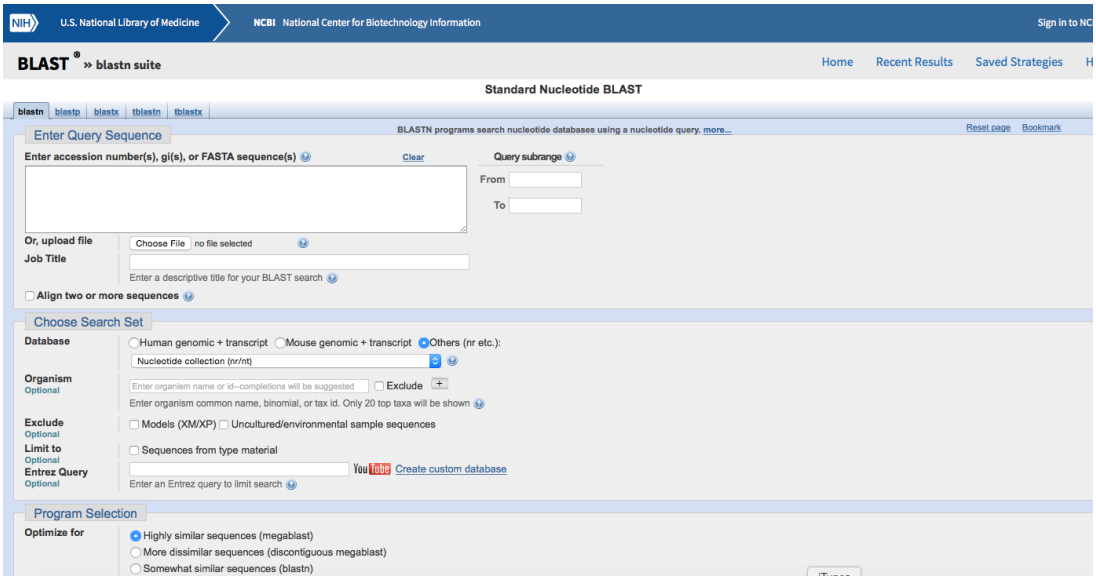
When two sequences are homologous, their amino acid or nucleotide sequences usually share significant **identity**. While 'homology' is a *qualitative* term; sequences are homologous or not - 'identity' and 'similarity' are *quantitative* that describe the degree of relatedness of sequences. Identity is the extent to which two sequences are exactly alike. Percent similarity of two sequences is the sum of both identical and similar residues. We infer homology when two sequences or structures share more similarity than would be expected by chance; when excess similarity is observed, the simplest explanation for that excess is that the two sequences did not arise independently, they arose from a common ancestor. Common ancestry explains excess similarity and therefore excess similarity implies common ancestry. However, sequence homology does not necessarily indicate functional homology. Phylogenetic homology does not necessarily imply structural homology or neither of them necessarily implies functional homology.

Sequence alignments are intended to unravel evolutionary pathways and/ or structural homology between two proteins. Relatedness of any two sequences can be assessed by performing a **pairwise alignment.** An alignment is called **'local'** when only a small subset of the two sequences is aligned. A **'global'** pairwise alignment includes end to end full length alignment of all residues of both sequences.

In a pairwise alignment, two sequences are directly compared next to each other to achieve maximal levels of identity (and maximal levels of conservation in the case of amino acid alignments). The purpose of a pairwise alignment is to assess the degree of similarity and the possibility of homology between two molecules. If the amount of sequence identity is sufficient, then the two sequences are probably homologous. Basic local alignment search tool (BLAST), FASTA, SSEARCH are commonly used similarity searching programs that provide accurate statistical estimates to reliably infer homology.

BLAST, SSEARCH, FASTA, and HMMER calculate local sequence alignments; local alignments identify the most similar region between two sequences. Searches with protein

sequences BLASTP, FASTP, SSEARCH or translated DNA sequences BLASTX, FASTX are preferred because they are many times more sensitive than DNA to DNA sequence comparison.



Picture 4: Standard Nucleotide BLAST - blastn webpage at NCBI

Proteins that are homologous may be **orthologous** or **paralogous. Orthologs** are homologous sequences in different species that arose from a common ancestral gene during speciation. thus in orthology history of sequence or gene reflects history of species. Paralogs are homologs that arose by mechanism of gene duplication without being followed by any speciation event. We thus define homologous sequences within the same organism as paralogous.

Trypsin and chymotrypsin are paralogous serine proteases with slightly different substrate specificities. Human alpha and human beta haemoglobin genes are paralogs while mouse alpha and human alpha hemoglobin is example of orthologs (Fig. 5).

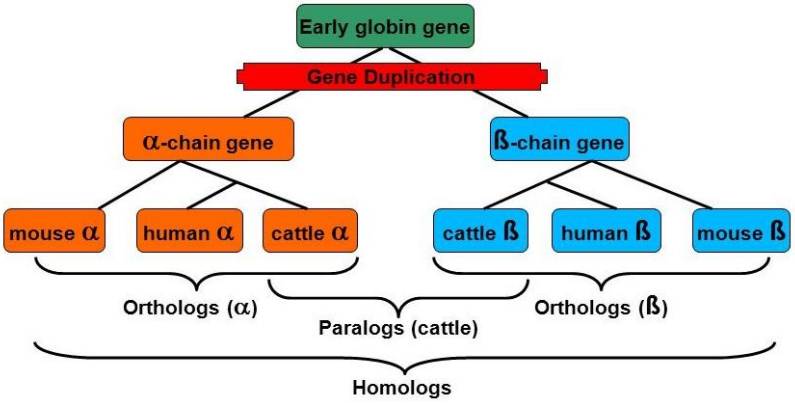


Fig. 5: Depicting homologous relationship between globin genes
http://images.slideplayer.com/16/4909104/slides/slide_4.jpg

Knowledge of molecular phylogenies and orthology in particular has become an integral component of many genome-scale studies of gene content, conserved gene order and gene expression, regulatory networks, metabolic pathways and in functional genome annotation Biologists are interested in identifying orthologs so as to find functionally equivalent genes (proteins) involved in a particular biological process (e.g. cell cycle) or metabolic pathway , to study fundamental processes and mechanisms of genome evolution (speciation, duplication or horizontal gene transfer), fate of genes and biological functions, or the genetic background of complex traits and inheritable diseases.

Databases such as Kyoto Encyclopedia of Genes and Genomes-KEGG, BioCyc, IMG integrate molecular data on pathways, enzymes and substrates associated with orthologous genes (proteins) from diverse genomes.

Table1: Popular Bioinformatic databases for ortholog determination

Bioinformatic databases for ortholog finding	URL's
KEGG-Kyoto Encyclopedia of Genes and Genomes	http://www.genome.jp/kegg/
IMG -Integrated Microbial Genomes	http://img.jgi.doe.gov/
BioCyc	http://biocy- c.org/
MBGD-Microbial Genome Database for comparative analysis	http://mbgd.genome.ad.jp/

The Clusters of Orthologous Groups (COGs) database has been designed to simplify evolutionary studies of complete genomes and improve functional assignments of individual proteins. Every COG contains orthologous sets of proteins from at least three phylogenetic lineages assumed to have evolved from an individual ancestral protein. Sequence similarity searches against the COG database can often suggest a possible function for a protein that otherwise has no clear database hits.

4. Summary

- Large amount of biological information is been generated from various research and ongoing sequencing projects.
- The need to simplify our understanding of biological complexity and to organise biological information as datasets has led to upsurge of bioinformatics as an applicative sub-discipline of biology.
- Bionformatics has emerged an interdisciplinary field based on concepts of various disciplines namely, biology, biophysics, biochemistry, computer science, mathematics and statistics.
- Foundations of bioinformatics were laid in a breakthrough work by Margaret Oakley Dayhoff through development of first comprehensive Atlas of protein sequence database in 1965 and establishment of one letter alphabetical code for identifying the amino acids.
- Single letter amino acid designation greatly paved way for easy storage of protein information in databases that would otherwise be very cumbersome for database handlings.
- Time since bioinformatics has seen tremendous development in terms of database creation, data management, data mining and advancements in software tools for accessing, searching, visualisation and retrieval of biological information.
- Bioinformatics deals with detection of genes, mRNA , proteins and metabolites; aiming primarily at holistic view of understanding and expression of macromolecules of life.

- These aims of bioinformatics can be achieved by way of three basic branches of bioinformatics namely; ‘genomics’ - ‘transcriptomics’ - ‘proteomics’ & epigenomics, pharmacogenomics , metabolomics, glycomics , lipidomics as their allied branches.

- Biological database a repository of biological data. Databases allow easy access, systemisation and analysis of computed biological information.

- Widespread applications of bioinformatics are in the field of forensic analysis, evolutionary studies, crop improvement, waste clean-up, phylogenetic analysis and comparative studies.

- Among key applications of bioinformatics include homologue finding - determining homologs, orthologs and paralogous sequences and drug designing.

- Discovery of novel drugs, molecular diagnostics and pharmacogenomics are some of the areas in which bioinformatics has become an integral part.

- BLAST, FASTA, SSEARCH are most popular pairwise sequence alignment tools for ascertaining homology. COGs, KEGG, IMB database are popularly used for ortholog finding and determination of functional homology among sequences. [1]