Concepts of Bioinformatics

1. Introduction

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. It is the emerging field that deals with the application of computers to the collection, organization, analysis, manipulation, presentation, and sharing of biologic data to solve biological problems on the molecular level. According to Frank Tekaia, bioinformatics is the mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information.

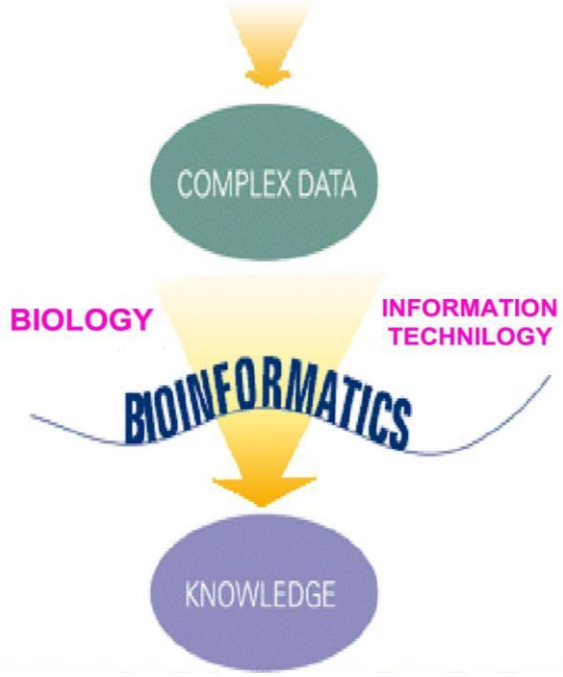


Fig 1. Concepts of Bioinformatics

The term *bioinformatics* was coined by Paulien Hogeweg in 1979 for the study of informatic processes in biotic systems. 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. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information."*

Bioinformatics is a scientific discipline that has emerged in response to accelerating demand for a flexible and intelligent means of storing, managing and querying large and complex biological data sets. The ultimate aim of bioinformatics is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. 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. At the beginning of the genomic revolution, the main concern of bioinformatics was the creation and maintenance of a database to store biological information such as nucleotide and amino acid sequences. Development of this type of database involved not only design issues but the development of an interface whereby researchers could both access existing data as well as submit new or revised data (e.g. to the NCBI, http://www.ncbi.nlm.nih.gov/). More recently, emphasis has shifted towards the analysis of large data sets, particularly those stored in different formats in different databases. Ultimately, all of this information must be combined to form a comprehensive picture of normal cellular activities so that researchers may study how these activities are altered in different disease states. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data, including nucleotide and amino acid sequences, protein domains, and protein structures.

2. Origin & History of Bioinformatics

Over a century ago, bioinformatics history started with an Austrian monk named Gregor Mendel. He is known as the "Father of Genetics". He cross-fertilized different colors of the same species of flowers. He kept careful records of the colors of flowers that he cross-fertilized and the color(s) of flowers they produced. Mendel illustrated that the inheritance of traits could be more easily explained if it was controlled by factors passed down from generation to generation.
After this discovery of Mendel, bioinformatics and genetic record keeping have come a long way. The understanding of genetics has advanced remarkably in the last thirty years. In 1972, Paul Berg made the first recombinant DNA molecule using ligase. In that same year, Stanley Cohen, Annie Chang and Herbert Boyer produced the first recombinant DNA organism. In 1973, two important things happened in the field of genomics:

1. Joseph Sambrook led a team that refined DNA electrophoresis using agarose gel, and
2. Herbert Boyer and Stanely Cohen invented DNA cloning. By 1977, a method for sequencing DNA was discovered and the first genetic engineering company, Genetech was founded.

During 1981, 579 human genes had been mapped and mapping by *in situ* hybridization had become a standard method. Marvin Carruthers and Leory Hood made a huge leap in bioinformatics when they invented a method for automated DNA sequencing. In 1988, the Human Genome Organization (HUGO) was founded. This is an international organization of scientists involved in Human Genome Project. In 1989, the first complete genome map was published of the bacteria *Haemophilus influenza*.

The following year, the Human Genome Project was started. In 1991, a total of 1879 human genes had been mapped. In 1993, Genethon, a human genome research center in France produced roduced a physical map of the human genome. Three years later, Genethon published the final version of the Human Genetic Map which concluded the end of the first phase of the Human Genome Project.

Bioinformatics was fuelled by the need to create huge databases, such as GenBank and EMBL and DNA Database of Japan to store and compare the DNA sequence data erupting from the human genome and other genome sequencing projects. Today, bioinformatics embraces protein structure analysis, gene and protein functional information, data from patients, pre-clinical and clinical trials, and the metabolic pathways of numerous species.

3. Importance

The greatest challenge facing the molecular biology community today is to make sense of the wealth of data that has been produced by the genome sequencing projects. Cells have a central core called nucleus, which is storehouse of an important molecule known as DNA. They are packaged in units known as chromosomes. They are together known as the genome. Genes are specific regions of the genomes (about 1%) spread throughout the genome, sometimes contiguous, many times non-contiguous. RNAs similarly contains informations, their major purpose is to copy information from DNA selectively and to bring it out of the nucleus for its use. Proteins are made of amino acids, which are twenty in count (researchers are debating on increasing this count, as couple of new ones are claimed to be identified).

The gene regions of the DNA in the nucleus of the cell is copied (transcribed) into the RNA and RNA travels to protein production sites and is translated into proteins is the Central Dogma of Molecular Biology. Portions of DNA Sequence are transcribed into RNA. The first step of a cell is to copy a particular portion of its DNA nucleotide sequence (i.e. gene) which is shown in Fig 2 and Fig 3.

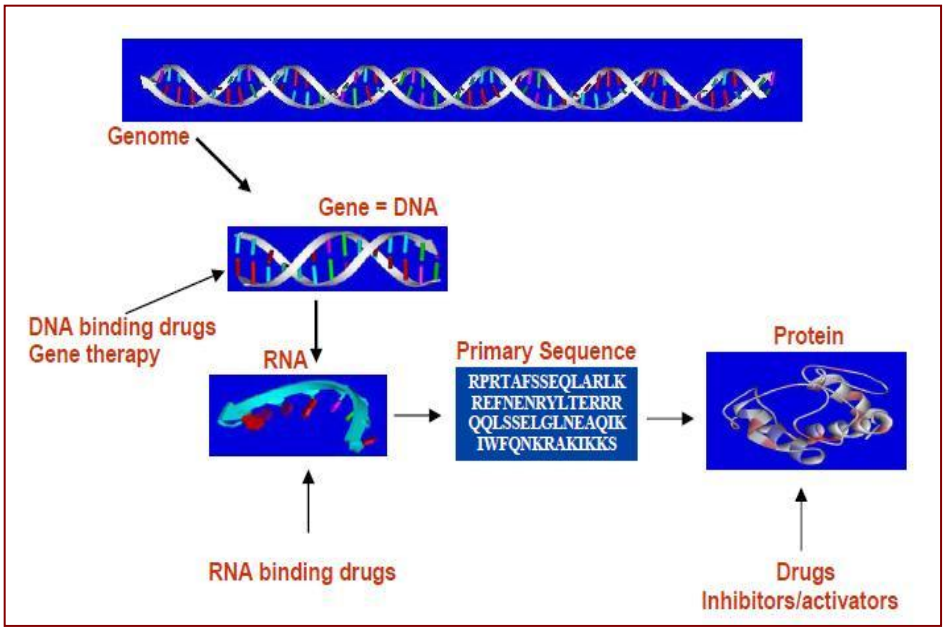


Fig 2. Biological Systems

Bioinformatics, being an interface between modern biology and informatics it involves discovery, development and implementation of computational algorithms and software tools that facilitate an understanding of the biological processes (Fig 3.) with the goal to serve primarily agriculture and healthcare sectors with several spinoffs. In a developing country like India, bioinformatics has a key role to play in areas like agriculture where it can be used for increasing the nutritional content, increasing the volume of the agricultural produce and implanting disease resistance etc. In the pharmaceutical sector, it can be used to reduce the time and cost involved in drug discovery process particularly for third world diseases, to custom design drugs and to develop personalized medicine.

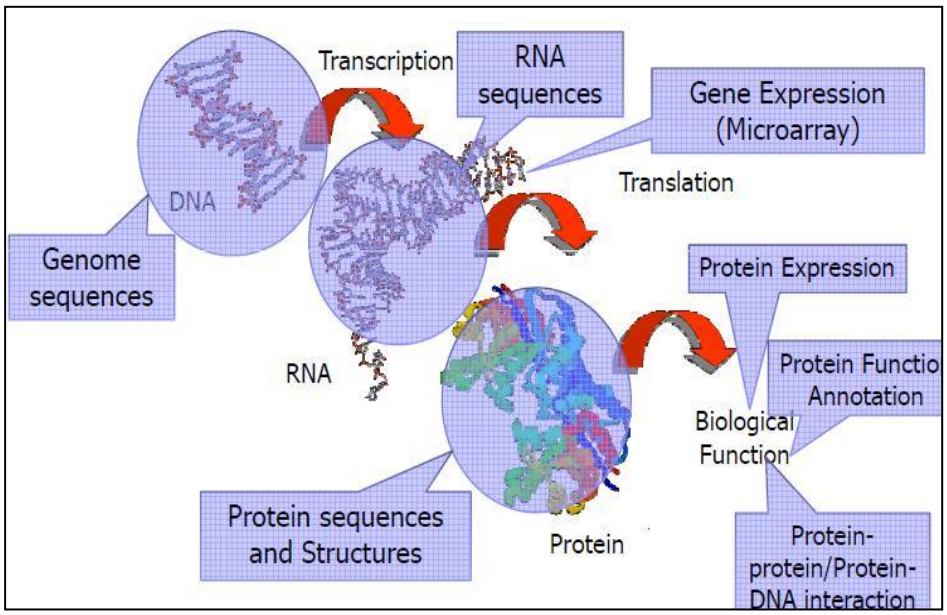


Fig 3. Biological Processes

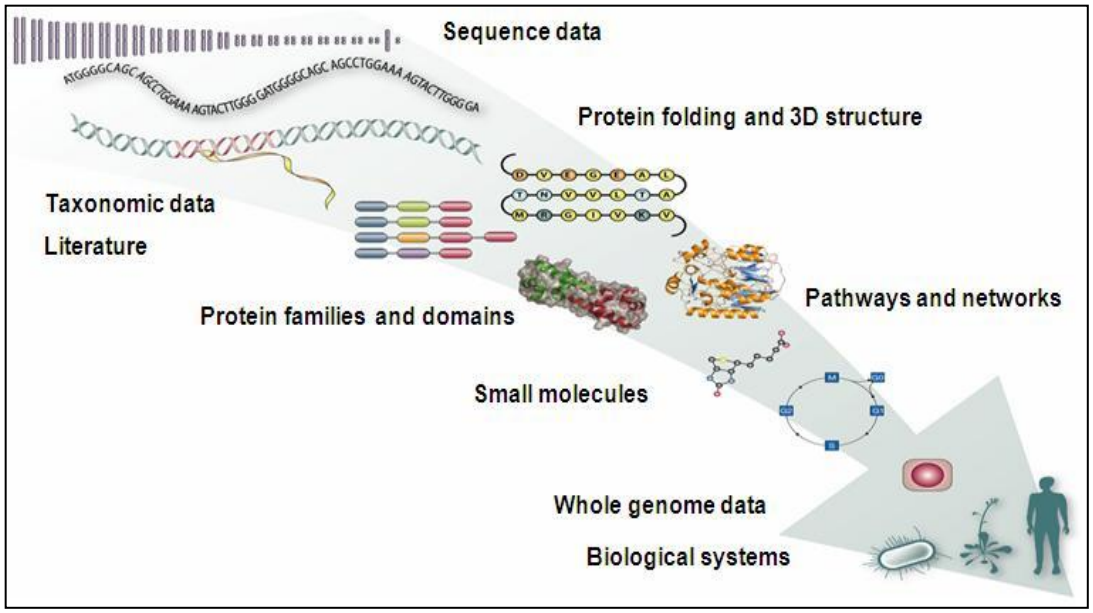
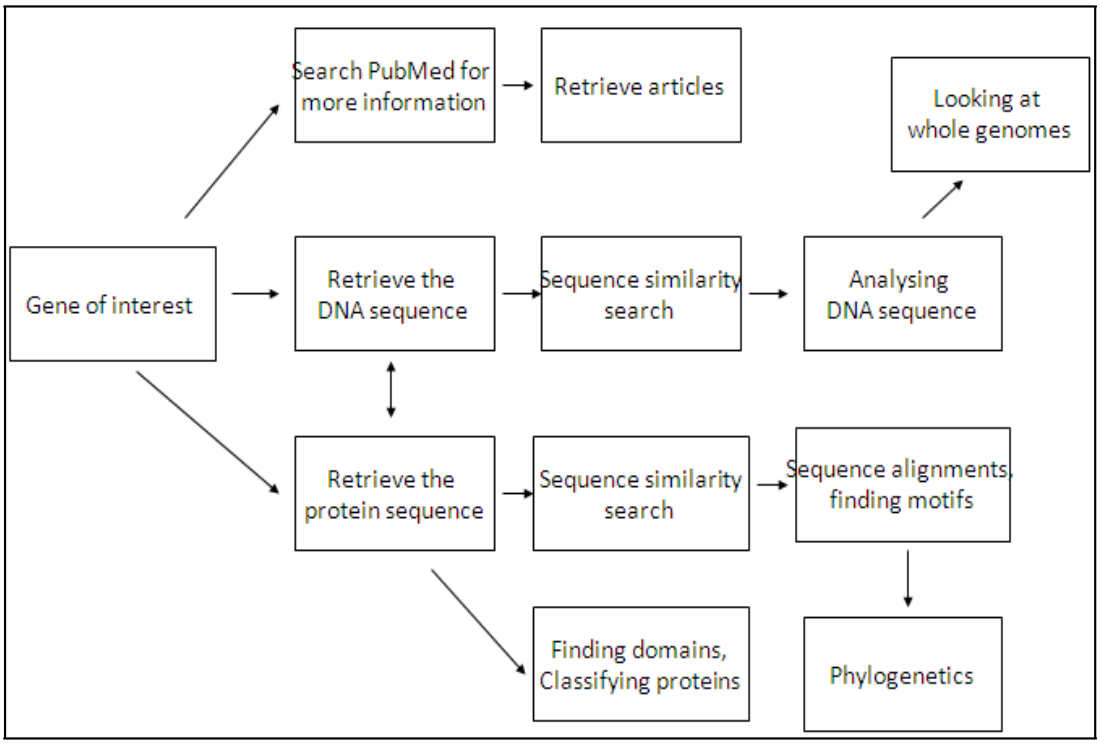


Fig 4. Information on Sequence Data

Traditionally, molecular biology research was carried out entirely at the experimental laboratory bench but the huge increase in the scale of data being produced in this genomic era has seen a need to incorporate computers into this research process. Sequence generation, its subsequent storage, interpretation and analysis are entirely computer dependent tasks. However, the molecular biology of an organism is a very complex issue with research being carried out at molecular level. The first challenge facing the bioinformatics community today is the intelligent and efficient storage of this massive data. Moreover, it is essential to provide easy and reliable access to this data. The data itself is meaningless before analysis and it is impossible for even a trained biologist to begin to interpret it manually. Therefore, automated computer tools must be developed to allow the extraction of meaningful biological information. There are three central biological processes around which bioinformatics tools must be developed:

- DNA sequence which determines protein sequence
- Protein sequence which determines protein structure
- Protein structure which determines protein function



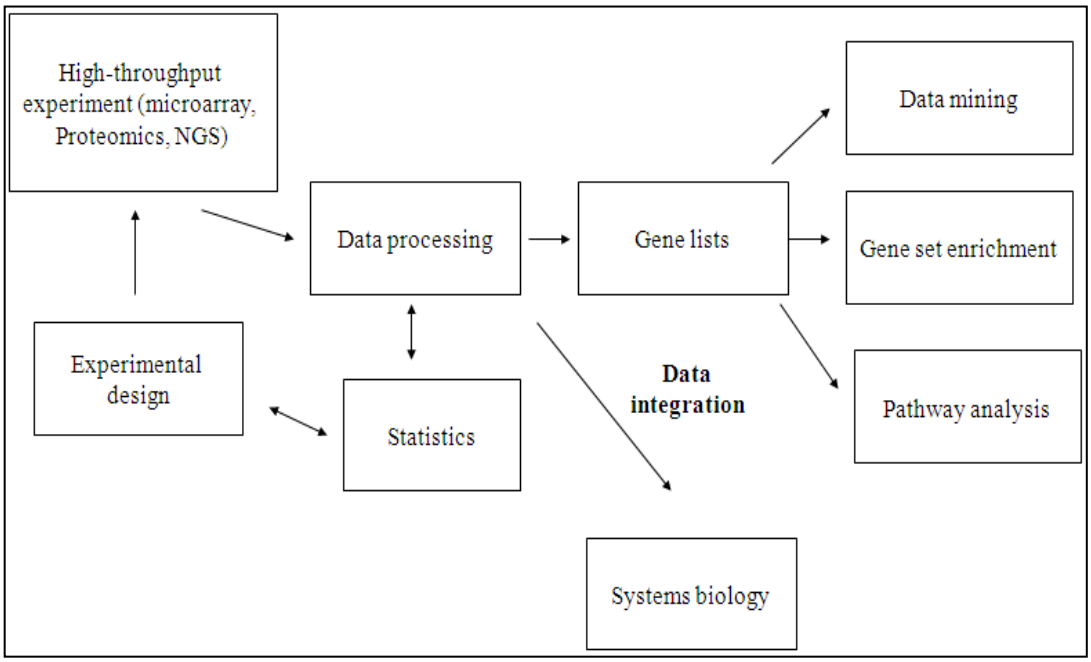


Fig. 5. Hypothesis-generating bioinformatics

4. Difference between Bioinformatics and Computational Biology

Both Bioinformatics and Computational Biology are Computers and Biology. Biologists who specialize in use of computational tools and systems to answer problems of biology are bioinformaticians. Computer scientists, mathematicians, statisticians, and engineers who specialize in developing theories, algorithms and techniques for such tools and systems are computational biologists. The actual process of analyzing and interpreting data is referred to as computational biology. Important sub- disciplines within bioinformatics and computational biology include:

- the development and implementation of tools that enable efficient access to, and use and management of, various types of information.
- the development of new algorithms (mathematical formulas) and statistics with which to assess relationships among members of large data sets, such as methods to locate a gene within a sequence, predict protein structure and/or function, and cluster protein sequences into families of related sequences

Bioinformatics has become a mainstay of genomics, proteomics, and all other *.omics (such as phenomics) and many information technology companies have entered the business or are considering entering the business, creating an IT (information technology) and BT (biotechnology) convergence. **A bioinformaticist** is an expert who not only knows how to use bioinformatics tools, but also knows how to write interfaces for effective use of the tools. **A bioinformatician**, on the other hand, is a trained individual who only knows to use bioinformatics tools without a deeper understanding.

5. Biological databases

Biological databases are huge data bases of mostly sequence data pouring in from many genome sequencing projects going on all over the world. They are an important tool in assisting scientists to understand and explain a host of biological phenomena from the structure of biomolecules and their interaction, to the whole metabolism of organisms to understanding the evolution of species. This knowledge helps facilitate to fight against diseases, assists in the development of medications and in discovering basic relationships amongst species in the history of life.

The information about DNA, proteins and the function of proteins must be stored in an intelligent fashion, so that scientists can solve problems quickly and easily using all available information. Therefore, the information is stored in *databanks*, many of which are accessible to everyone on the internet. A few examples are a databank containing protein structures (the PDB or Protein Data Bank), a databank containing protein sequences and their function (Swiss-Prot), a databank with information about enzymes and their function (ENZYME), and a databank with nucleotide sequences of all genes sequenced up to date (EMBL). Due to the current state of technology, there are large differences between the sizes of databanks. EMBL, the nucleotides database contains many more sequences than the number of protein structures registered in the PDB. The reason for this is that it is a lot simpler to sequence a gene, than to find out which protein is encoded by this gene and what its function is. Also it is more difficult to determine the structure of the protein.

Using databanks, one can perform all kinds of comparisons and search queries. If, for example, you know a protein which causes a disease in humans, your might look into a databank to

see if a similar protein has previously been described and what this protein does in the human body.

Using known information will make it easier and quicker to develop a drug against the disease or a test to detect the disorder in an early stage.

The Biological data can be broadly classified as:

Biological Databases	Information they contain
1. Bibliographic databases	Literature
2. Taxonomic databases	Classification
3. Nucleic acid databases	DNA information
4. Genomic databases	Gene level information
5. Protein databases	Protein information
6. Protein families, domains and functional sites	Classification of proteins and identifying domains
7. Enzymes/ metabolic pathways	Metabolic pathways

There are many different types of database but for routine sequence analysis, the following are initially the most important

1. Primary databases: Contain sequence data such as nucleic acid or protein. Example of primary databases include:

Protein Databases	Nucleic Acid Databases
• SWISS-PROT	• EMBL
• TREMBL	• Genbank
• PIR	• DDBJ

2. Secondary databases: These are also known as pattern databases contain results from the analysis of the sequences in the primary databases. Example of secondary databases include: PROSITE, Pfam, BLOCKS, PRINTS.

6. Introduction to NCBI and Entrez

The web-site of National Center for Biotechnology Information (NCBI) is one of the world's premier website for biomedical and bioinformatics research (http://www.ncbi.nlm.nih.gov/). Based within the National Library of Medicine at the National Institutes of Health, USA, the NCBI hosts many databases used by biomedical and research professionals. The services include PubMed (the bibliographic database); GenBank (the nucleotide sequence database); and the BLAST algorithm for sequence comparison, among many others. It is established in 1988 as a national resource for molecular biology information. NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information all for the better understanding of molecular processes affecting human health and disease.

Every database has a unique identifier. Each entry in a database must have a unique identifier EMBL Identifier (ID), GENBANK Accession Number (AC). This database stores information along with the sequence. Each piece of information is written on it's own line, with a code defining the line. For example, DE (description); OS (organism species); AC (accession number). Relevant biological information is usually described in the feature table (FT).

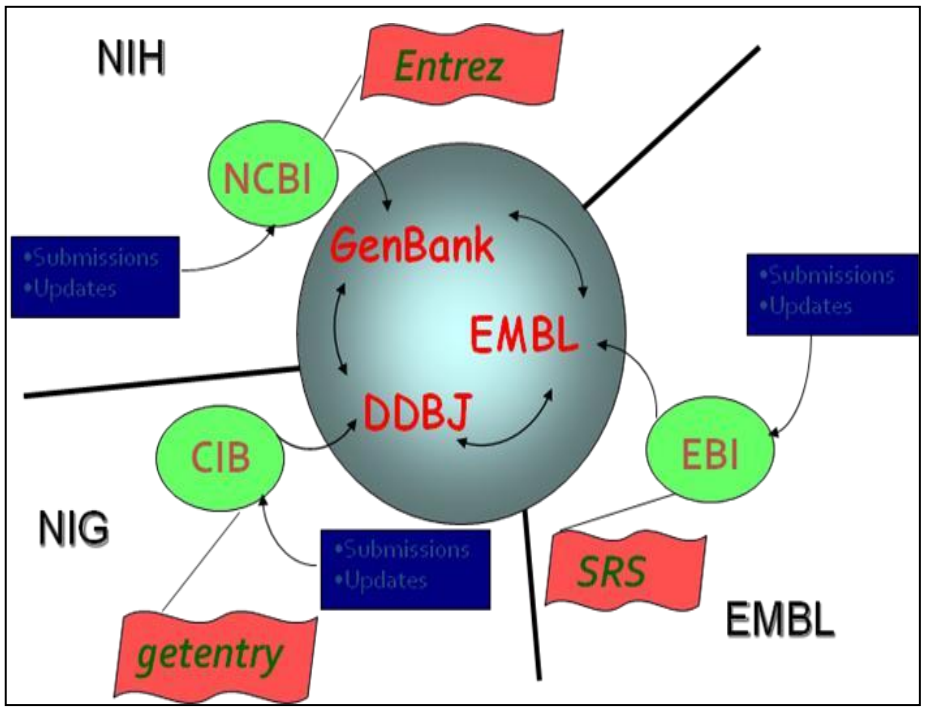


Fig 6. International Sequence Database Collaboration

7. The Entrez Search and Retrieval System

Entrez is the text-based search and retrieval system used at NCBI for all of the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, OMIM, and many others. Entrez is at once an indexing and retrieval system, a collection of data from many sources, and an organizing principle for biomedical information. These general concepts are the focus of this section (Fig 7.).

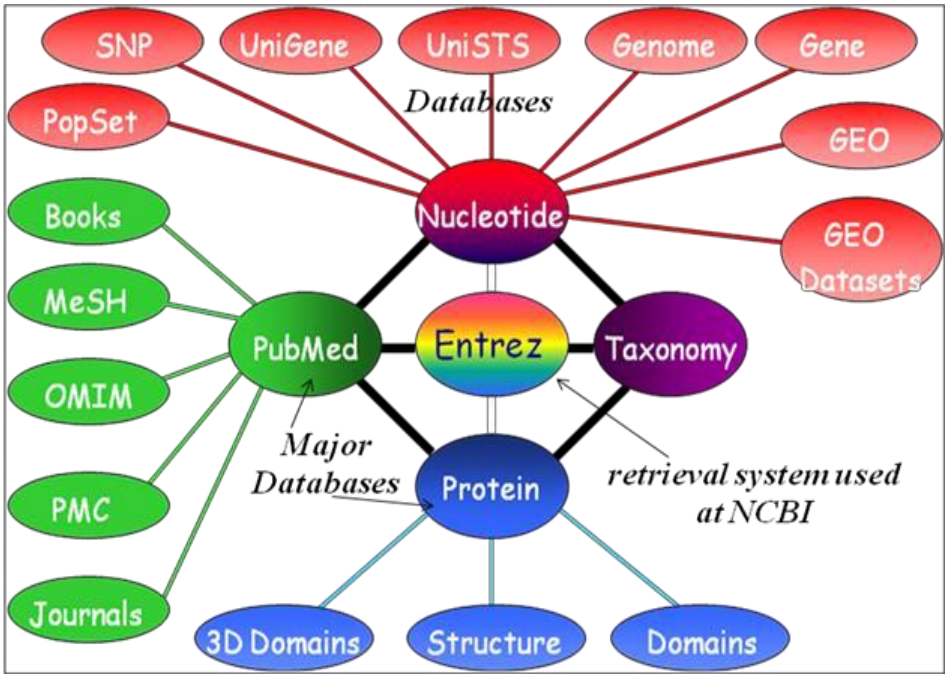


Fig 7. NCBI - RDBMS

8. The Nucleotide Sequence database

The GenBank sequence database is an annotated collection of all publicly available nucleotide sequences and their protein translations. This database is produced at National Center for Biotechnology Information (NCBI) as part of an international collaboration with the European Molecular Biology Laboratory (EMBL) as given in Fig. 8, data library from the European Bioinformatics Institute (EBI) and the DNA Data Bank of Japan (DDBJ) given in Fig. 9. GenBank and its collaborators receive sequences produced in laboratories throughout the world from more than 100,000 distinct organisms. GenBank continues to grow at an exponential rate, doubling every 10 months. Release 134, produced in February 2003, and contained over 29.3 billion nucleotide bases in more than 23.0 million sequences. GenBank is built by direct submissions from individual laboratories, as well as from bulk submissions from large-scale sequencing centers.

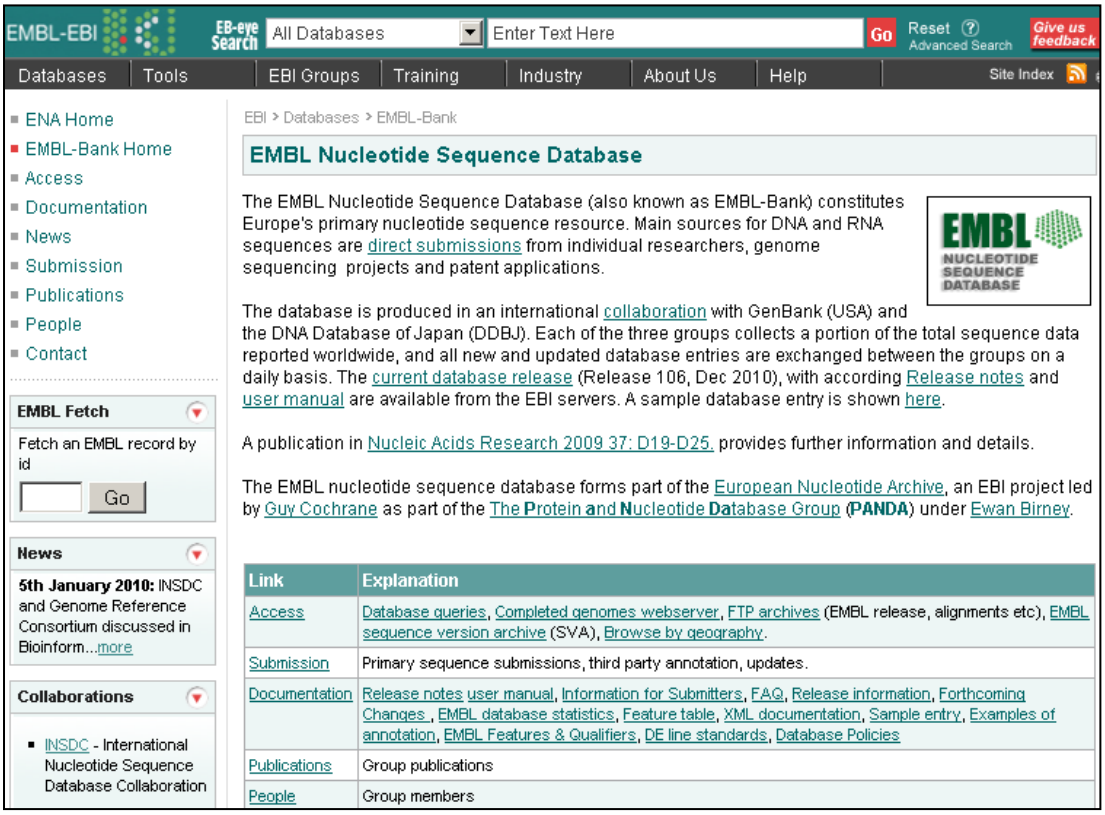


Fig. 8 EMBL Nucleotide Sequence Database

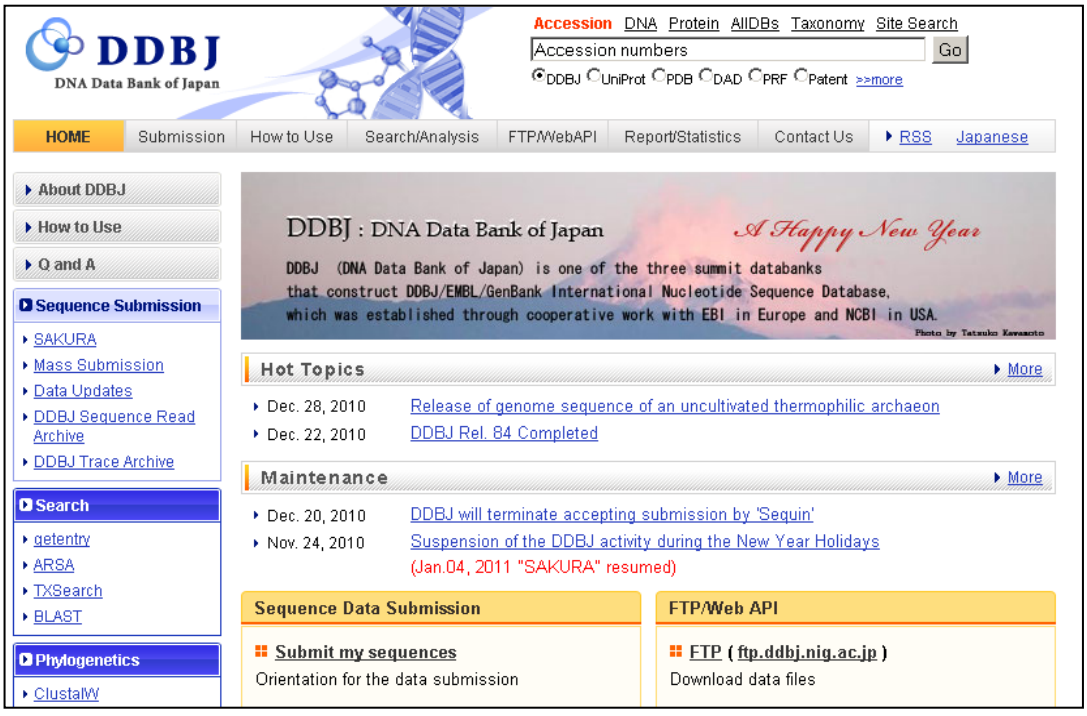


Fig. 9. DNA Data Bank of Japan

9. The Bibliographic Database

PubMed is a database developed by the NCBI. The database was designed to provide access to citations (with abstracts) from biomedical journals. Subsequently, a linking feature was added to provide access to full-text journal articles at Web sites of participating publishers, as well as to other related Web resources. PubMed is the bibliographic component of the NCBI's Entrez retrieval system.

MEDLINE is NLM's premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, and the preclinical sciences. Journal articles are indexed for MEDLINE, and their citations are searchable, using NLM's controlled vocabulary, MeSH (Medical Subject Headings). MEDLINE contains all citations published in Index Medicus, and corresponds in part to the International Nursing Index and the Index to Dental Literature.

10. Macromolecular Structure Databases

The resources provided by NCBI for studying the three-dimensional (3D) structures of proteins center around two databases: the Molecular Modeling Database (MMDB), which provides structural information about individual proteins; and the Conserved Domain Database (CDD), which provides a directory of sequence and structure alignments representing conserved functional domains within proteins(CDs). Together, these two databases allow scientists to retrieve and view structures, find structurally similar proteins to a protein of interest, and identify conserved functional sites.

11. Computer Programming in Bioinformatics: JAVA in Bioinformatics

The geographical scattered research centres all around the globe ranging from private to academic settings, and a range of hardware and OSs are being used, Java is emerging as a key player in bioinformatics. Physiome Sciences' computer-based biological simulation technologies and Bioinformatics Solutions' PatternHunter are two examples of the growing adoption of Java in bioinformatics.

12. Perl in Bioinformatics

String manipulation, regular expression matching, file parsing, data format interconversion etc are the common text-processing tasks performed in bioinformatics. Perl excels in such tasks and is being used by many developers. Yet, there are no standard modules designed in Perl specifically for the field of bioinformatics. However, developers normally designed several of their own individual modules for any specific purpose, which have become quite popular and are coordinated by the BioPerl project.

13. Measuring biodiversity

Biodiversity Databases are used to collect the species names, descriptions, distributions, genetic information, status & size of populations, habitat needs, and how each organism interacts with other species etc. Computer simulations models are useful to study population dynamics, or calculate the cumulative genetic health of a breeding pool (in agriculture) or endangered population (in conservation). Entire DNA sequences or genomes of endangered species can be preserved, allowing the results of Nature's genetic experiment to be remembered *in silico*.

In these days of growing human population and habitat destruction, knowledge of centers of high biodiversity is critical for rational conservation decisions to be made. The major problem area is that this information is largely unavailable to the decision makers. It is ironic that most of these data are in the great museums, which are located in the cool temperate parts of the world whereas; most of the organisms are in the warm humid parts of the world. The data that exist are paper based. Descriptions by collectors and curators, herbarium sheets, diagrams and photographs, and of course, pickled and preserved specimens with their labels. If a researcher wishes to consult these data he/she has to travel to the museum in question. For people who need a breadth of information to make decisions, this is obviously not an option. There are two areas in biology where enormous amounts of information are generated. One is in molecular biology which deals with base sequences in DNA and amino acid sequences in proteins, and the other is the biodiversity information crisis. Mathematics and computers are being used to tackle these problems with procedures which come under the label of Bioinformatics.

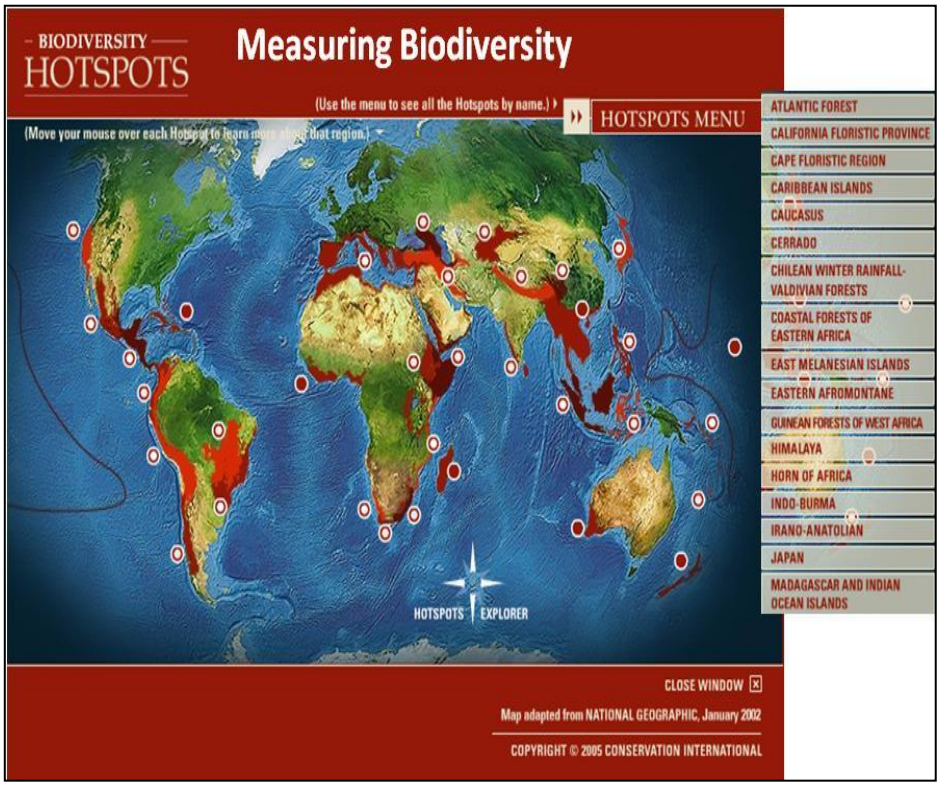


Fig 10. Biodiversity Hotspots regions

14. Sequence analysis and alignment

The most well-known application of bioinformatics is sequence analysis. In sequence analysis, DNA sequences of various organisms are stored in databases for easy retrieval and comparison. The well-reported Human Genome Project (Fig. 11) is an example of sequence analysis bioinformatics. Using massive computers and various methods of collecting sequences, the entire human genome was sequenced and stored within a structured database. DNA sequences used for bioinformatics can be collected in a number of ways. One method is to go through a genome and search out individual sequences to record and store. Another method is to compare all fragments for finding whole sequences by overlapping the redundant segments. The latter method, known as shotgun sequencing, is currently the most popular because of its ease and speed. By comparing known sequences of a genome to specific mutations, much information can be assembled about undesirable mutations such as cancers. With the completed mapping of the human genome, bioinformatics has become very important in the research of cancers in the hope of an eventual cure. Computers are also used to collect and store broader data about species. The Species 2000 project, for example, aims to collect a large amount of information about every species of plant, fungus, and animal on the earth. This information can then be used for a number of applications, including tracking changes in populations and biomes.

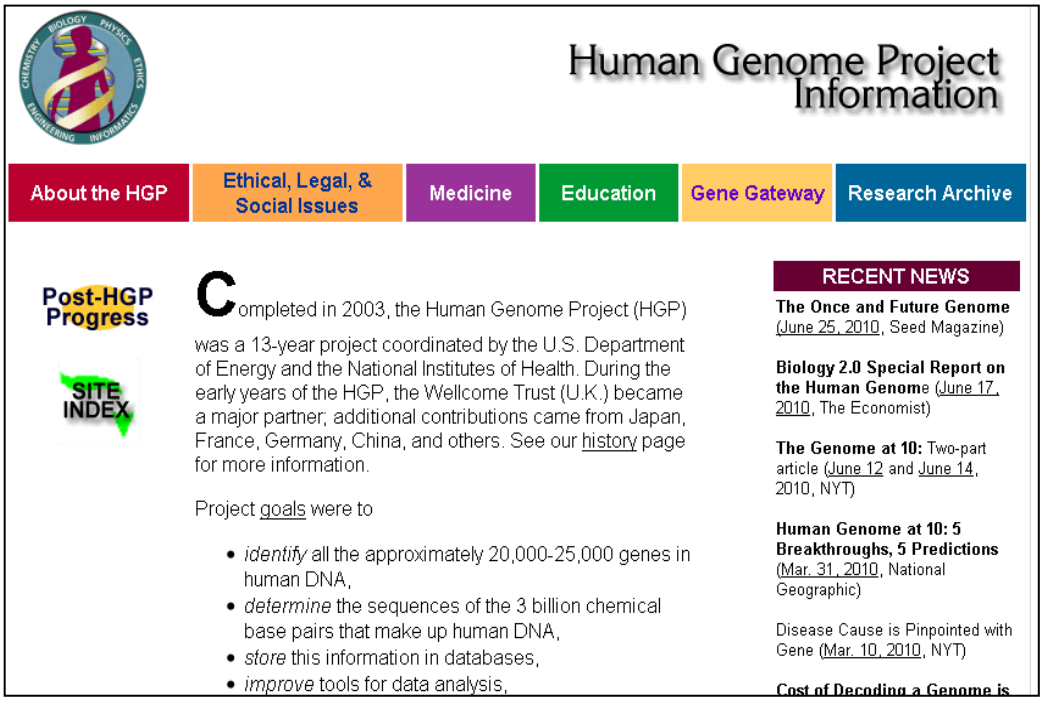


Fig. 11. Human Genome Project

With the growing amount of data, earlier it was impractical to analyze DNA sequences manually. Nowadays, many tools and techniques are available provide the sequence comparisons (sequence alignment) and analyze the alignment product to understand the biology. For example, BLAST is used to search the genomes of thousands of organisms, containing billions of nucleotides. BLAST is software which can do this using dynamic programming, as fast as google searches for your keywords, considering the length of query words of bio-sequences.

Sequence Alignment: The sequence alignment can be categorized into two groups i.e. global and local alignment

Global Alignment
Input: two sequences S_1, S_2 over the same alphabet
Output: two sequences S'_1, S'_2 of equal length
(S'_1, S'_2 are S_1, S_2 with possibly additional gaps)
Example:

- S_1= GCGCATGGATTGAGCGA
- S_2= TGCGCCATTGATGACC
- A possible alignment:

S'_1= -GCGC-ATGGATTGAGCGA
S'_2= TGCGCCATTGAT-GACC—

Local Alignment
Goal: Find the pair of substrings in two input sequences which have the highest similarity
Input: two sequences S_1, S_2 over the same alphabet
Output: two sequences S''_1, S''_2 of equal length
(S'_1, S'_2 are substrings of S_1, S_2 with possibly additional gaps)
Example:

- S_1= GCGCATGGATTGAGCGA
- S_2= TGCGCCATTGATGACC
- A possible alignment:

S'_1= ATTGA-G
S'_2= ATTGATG

FASTA: In bioinformatics, FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which base pairs or amino acids are represented using single-letter codes. The format also allows for sequence names and comments to precede the sequences. The FASTA format may be used to represent either single sequences or many sequences in a single file. A series of single sequences, concatenated, constitute a multisequence file. A sequence in FASTA format is represented as a series of lines, which should be no longer than 120 characters and usually do not exceed 80 characters. This probably was because to allow for preallocation of fixed line sizes in software: at the time, most users relied on DEC VT (or compatible) terminals which could display 80 or 132 characters per line. Most people would prefer normally the bigger font in 80-character modes and so it became the recommended fashion to use 80 characters or less (often 70) in FASTA lines. The first line in a FASTA file starts either with a ">" (greater-than) symbol or a ";" (semicolon) and was taken as a comment. Subsequent lines starting with a semicolon would be ignored by software. Since the only comment used was the first, it quickly became used to hold a summary description of the sequence, often starting with a unique library accession number, and with time it has become commonplace use to always use ">" for the first line and to not use ";" comments (which would otherwise be ignored).

>gi|5524211|gb|AAD44166.1| cytochrome b [*Elephas maximus maximus*]

LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSA
IPYIGTNLVEWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDS
DKIPFHPYYTIKDFLGLLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWFLFAYAI
LRSVPNKLGGVLALFLSIVILGLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQP
VEYPYTIIGQMASILYFSIILAFLPIAGXIENY

15. Prediction of protein structure

Proteins play crucial functional roles in all biological processes: enzymatic catalysis, signaling messengers, structural elements. Function depends on unique 3-D structure. It is easy to obtain protein sequences but difficult to determine structure. Protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. In the vast majority of cases, this primary structure uniquely determines a structure in its native environment. Knowledge of this structure is vital in understanding the function of the protein. For lack of better terms, structural information is usually classified as one of *secondary*, *tertiary* and *quaternary* structure. Protein structure prediction is the prediction of the three-dimensional structure of a protein from its amino acid sequence i.e, the prediction of its tertiary structure from its primary structure. Protein structures are being determined with increasing speed. Consequently, automated and fast bioinformatics tools are required for exploring structure–function relationships in large numbers of proteins. These are necessary both when the function has been characterized experimentally and when it must be predicted.

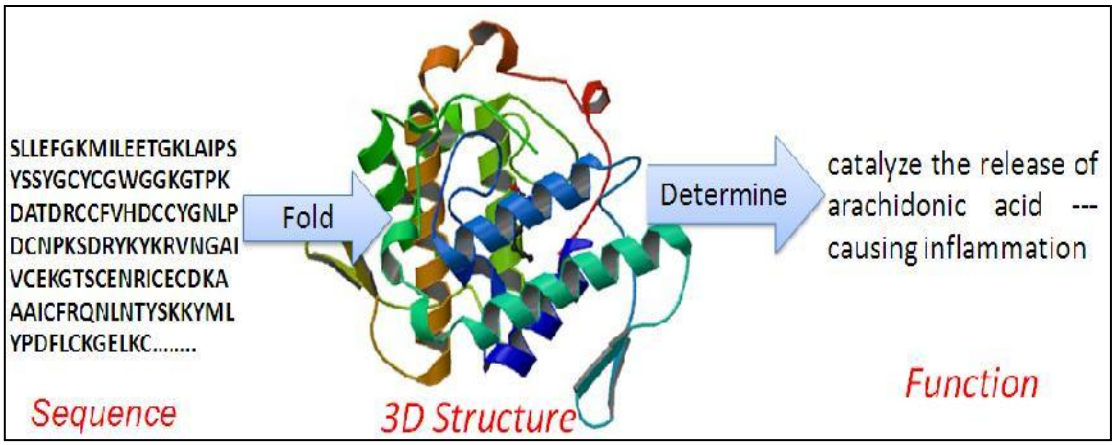


Fig. 12. Protein Structure Prediction

In the genomic branch of bioinformatics, homology is used to predict the function of a gene: if the sequence of gene *A*, whose function is known, is homologous to the sequence of gene *B*, whose function is unknown, one could infer that B may share A's function. In the structural branch of bioinformatics, homology is used to determine which parts of a protein are important in structure formation and interaction with other proteins. In a technique called homology modeling, this information is used to predict the structure of a protein once the structure of a homologous protein is known. One example of this is the similar protein homology between hemoglobin in humans and the hemoglobin in legumes (leghemoglobin). Both serve the same purpose of transporting oxygen in the

organism. Though both of these proteins have completely different amino acid sequences, their protein structures are virtually identical, which reflects their near identical purposes.

16. Molecular docking

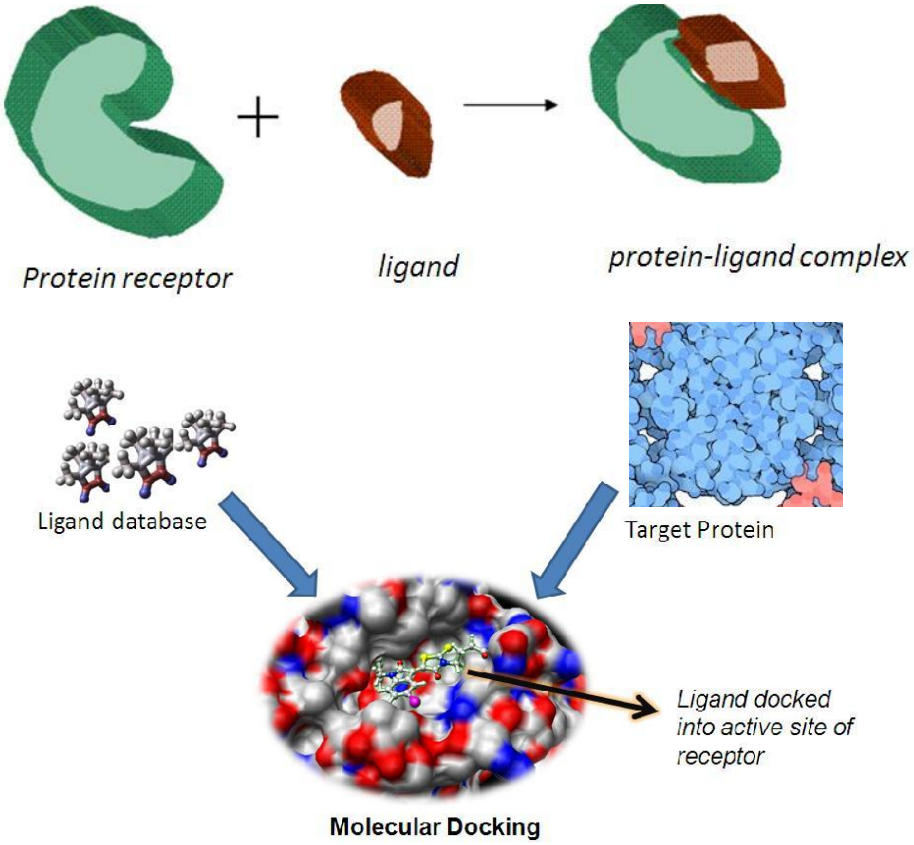


Fig. 13 Protein-ligand Docking

In the last two decades, tens of thousands of protein three-dimensional structures have been determined by X-ray crystallography and Protein nuclear magnetic resonance spectroscopy (protein NMR). One central question for the biological scientist is whether it is practical to predict possible protein-protein interactions only based on these 3D shapes, without doing protein-protein interaction experiments. A variety of methods have been developed to tackle the Protein-protein docking problem, though it seems that there is still much work to be done in this field. We are interested in information about our DNA, proteins and the function of proteins. Genes and proteins can be sequenced, so the sequence of bases in genes or amino acids in proteins can be determined. This information must be stored in an intelligent fashion, so that scientists can solve problems quickly and easily using all available information. Therefore, the information is stored in *databanks*, many of which are accessible to everyone on the internet. A few examples are a databank containing protein structures (the PDB or Protein Data Bank), a databank containing protein sequences and their function (Swiss-Prot), a databank with information about enzymes and their function (ENZYME), and a databank with nucleotide sequences of all genes sequenced up to date (EMBL).

17. Bioinformatics in Agriculture

The most critical tasks in bioinformatics involves the finding of genes in the DNA sequences of various organisms, developing methods to predict the structure and function of newly discovered proteins and structural RNA sequences, clustering protein sequences into families of related sequences, development of protein models, aligning similar proteins and generating phylogenetic trees to examine evolutionary relationships. The sequencing of the genomes of microbes, plants and animals should have enormous benefits for the agricultural community. Computational analysis of these sequence data generated by genome sequencing, proteomics and array-based technologies is critically important. Bioinformatics tools can be used to search for the genes within these genomes and to elucidate their functions.

The sequencing of the genomes of plants and animals should have enormous benefits for the agricultural community. Bioinformatic tools can be used to search for the genes within these genomes and to elucidate their functions. This specific genetic knowledge could then be used to produce stronger, more drought, disease and insect resistant crops and improve the quality of livestock making them healthier, more disease resistant and more productive.

18. Bioinformatics in India

Studies of IDC points out that India will be a potential star in bioscience field in the coming years after considering the factors like bio-diversity, human resources, infrastructure facilities and governments initiatives.

Bioinformatics has emerged out of the inputs from several different areas such as biology, biochemistry, biophysics, molecular biology, biostatics, and computer science. Specially designed algorithms and organized databases is the core of all informatics operations. The requirements for such an activity make heavy and high level demands on both the hardware and software capabilities. This sector is the quickest growing field in the country. The vertical growth is because of the linkages between IT and biotechnology, spurred by the human genome project. The promising start-ups are already there in Bangalore, Hyderabad, Pune, Chennai, and Delhi. There are over 200 companies functioning in these places. IT majors such as Intel, IBM, Wipro are getting into this segment spurred by the promises in technological developments.

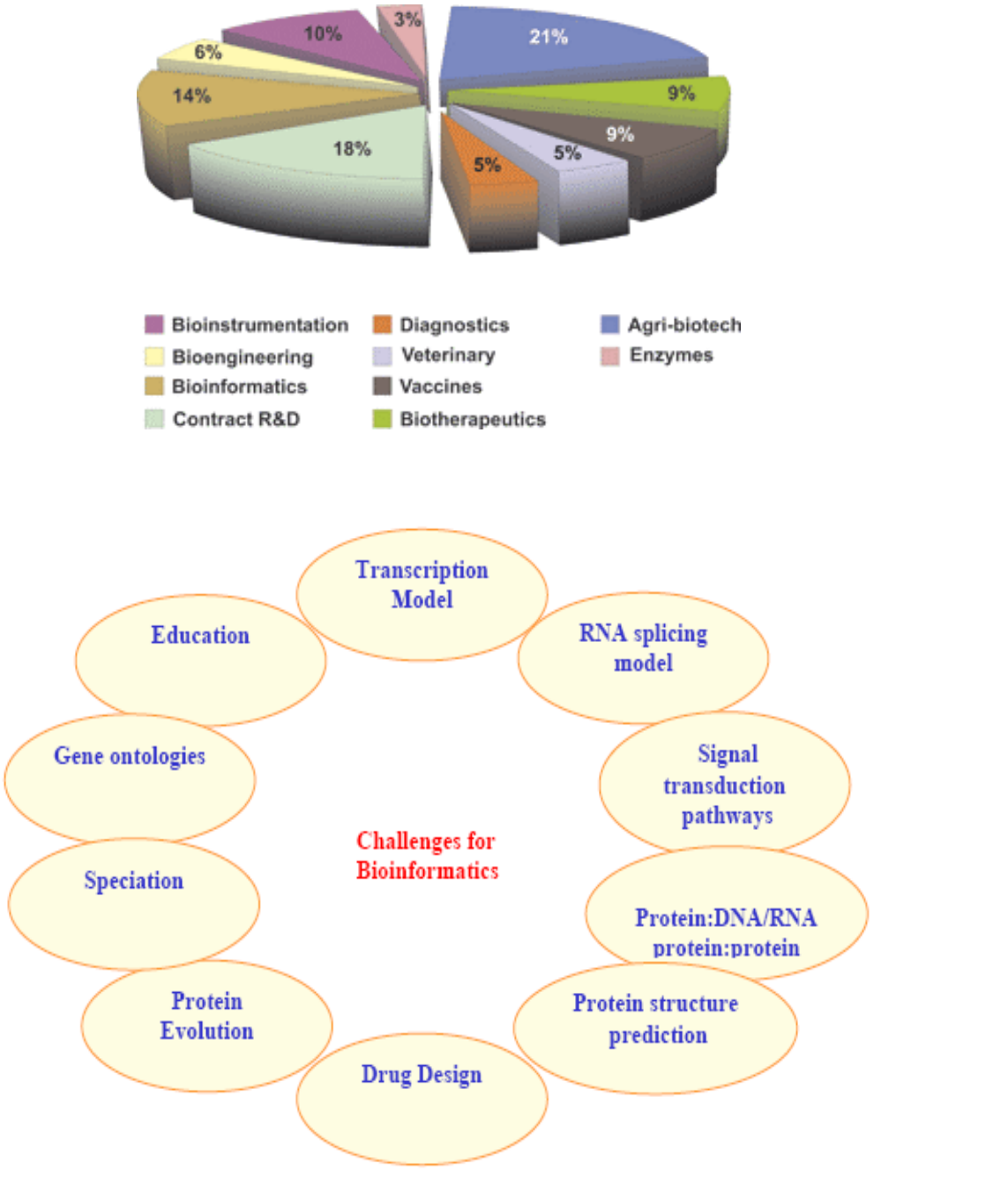


Fig. 14. Applications and Challenges in Bioinformatics

References

1. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Wheeler,D.L.(2005) GenBank. *Nucleic Acids Research*, 33, D34–D38.
2. Bioinformatics in the 21st century (1998). *A report to the research resources and Infrastructure working group subcommittee on biotechnology national science and Technology council white house office of science and technology policy Bioinformatics.*
3. Crick F. (1970). Central Dogma of Molecular Biology. *Nature*, 227, 561-563.
4. http://www.ncbi.nlm.nih.gov/books/nbk21101/
5. Human genome project and beyond (ww.ornl.gov/hgmis/)
6. Indigenous knowledge, Bioinformatics and Rural Agriculture. (2005). *9th ICABR International Conference on Agricultural Biotechnology: ten years later*, Ravello (Italy), July 6 to July 10.
7. Jayaram B and Priyanka D. *Bioinformatics for a better tomorrow*. Department of chemistry & Supercomputing facility for bioinformatics & computational biology, Indian Institute of Technology.
8. Maglott D., Ostell J., Pruitt K. D. and Tatusova T. (2005) Entrez gene: gene-centered information at NCBI, *Nucleic Acids Research*, 33, D54–D58.
9. Mcentyre J. and Ostell J. (2005). *The NCBI Handbook*. Bethesda (MD): National Library of Medicine (US)
10. McEntyre Jo, Jim O., National Center for Biotechnology Information Bethesda (MD): National Center for Biotechnology Information (US); 2002. *The NCBI Handbook.* Medicine (US), NCBI.
11. Ronald M. A., Knegtel, Irwin D. Kuntz and Oshiro C. M. (1997). Molecular Docking to Ensembles of Protein Structures. *Journal of Molecular. Biology.* 266, 424-440
12. Wheeler,D.L., Benson,D.A., Bryant,S., Canese,K., Church,D.M., Edgar,R., Federhen,S., Helmberg,W., Kenton,D., Khovayko,O. et al. (2005) Database resources of the National Center for Biotechnology Information: Update. *Nucleic Acid Research*, 33, D39–D45.