BHARATHIDASAN UNIVERSITY

Tiruchirappalli- 620024

Tamil Nadu, India.

Programme: M.Sc. Statistics

Course Title: Statistical Methods for Bioinformatics

Course Code: 23ST08DEC

Unit-I

Introduction to Bioinformatics

Dr. T. Jai Sankar
Associate Professor and Head
Department of Statistics

Ms. I. Angel Agnes Mary
Guest Faculty
Department of Statistics

UNIT - I

Introduction to Bioinformatics

Bioinformatics is a dynamic and emerging field of science which deals with an amalgamation of several subjects. These subjects include Biology, Chemistry, Mathematics, Statistics, and Computer Science. Bioinformatics focuses on developing new technologies in the fields of medicine, research, and biotechnology. This subject is interdisciplinary and requires thorough knowledge of both engineering as well as life sciences. This sector draws from a well of biological data and uses this information to create new tools and software which will be relevant in the world of biological research. In this article, we'll explore what is bioinformatics, the application of bioinformatics, the scope of bioinformatics, and the uses of bioinformatics.

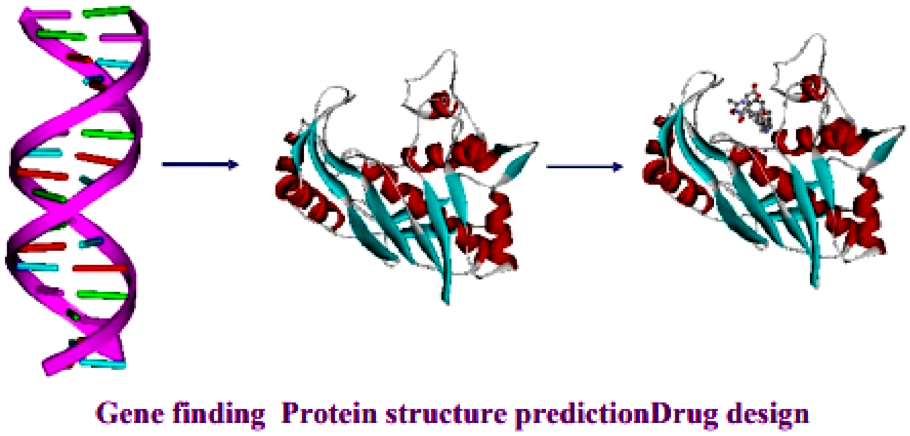


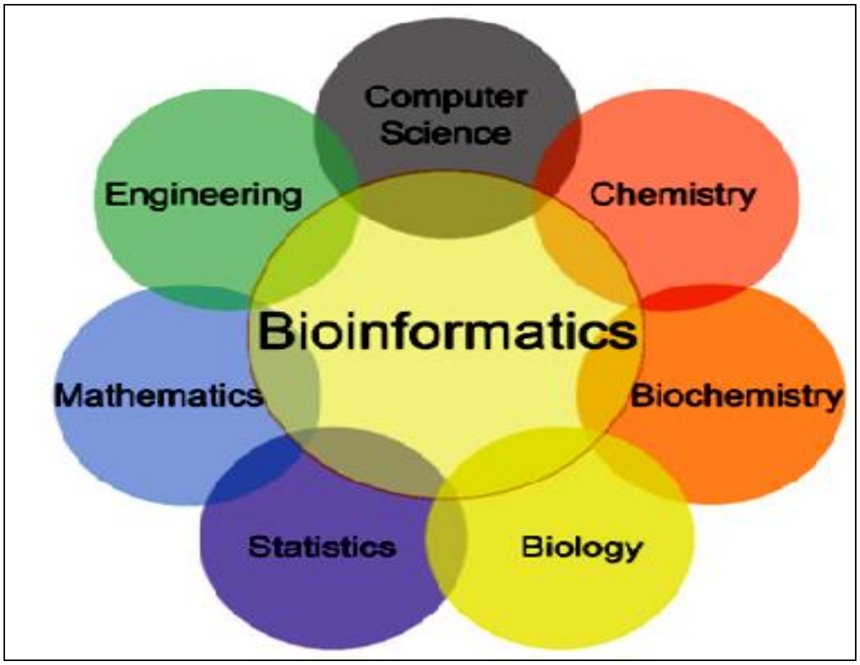
Figure 1: Some major areas of research in bioinformatics and computational biology.

This is the figure for Gene finding Protein prediction Drug design in some major areas of research in bioinformatics and computational biology.

Bioinformatics is the combination of biology and information technology. Basically, bioinformatics is a recently developed science using information to understand biological phenomenon. It broadly involves the computational tools and methods used to manage, analyse and manipulate volumes and volumes of biological data.

Bioinformatics deals with storage, retrieval, analysis and interpretation of biological data using computer based software and tools.

Bioinformatics is a branch of science that integrates computer science, mathematics and statistics, chemistry and engineering for analysis, exploration, integration and exploitation of biological sciences data, in Research and Development.



History of Bioinformatics

- Bioinformatics emerged in mid 1990s.
- From 1965-78 Margaret O. Dayhoff established first database of protein sequences, published annually as series of volume entitled “Atlas of protein sequence and structure”.
- During 1977 DNA sequences began to accumulate slowly in literature and it became more common to predict protein sequences by translating sequenced genes than by direct sequencing of proteins.
- Thus number of uncharacterized proteins began to increase.
- In 1980, there were enough DNA sequences to justify the establishment of the first nucleotide sequence database, GenBank at National Centre for Biotechnology Information (NCBI), USA. NCBI served as primary databank provider for information.
- The European Molecular Biology Laboratory (EMBL) established at European Bioinformatics Institute (EBI) in 1980. The aim of this data library was to collect, organize and distribute nucleotide sequence data and related information.
- In 1986 DNA Data Bank was established by GemonNet, Japan.
- In 1984, the National Biomedical Research Foundation (NBRF) established the protein information Resource (PIR).
- All these data banks operate in close collaboration and regularly exchange data.
- Management and analysis of the rapidly accumulating sequence data required new computer software and statistical tools.
- This attracted scientists from computer science and mathematics to the fast emerging field of bioinformatics.

Definition of Bioinformatics

- Roughly, bioinformatics describes any use of computers to handle biological information. In practice the definition used by most people is narrower; bioinformatics to them is a synonym for "computational molecular biology"- the use of computers to characterize the molecular components of living things.
- **"Classical" bioinformatics:** "The mathematical, statistical and computing methods that aim to solve biological problems using DNA and amino acid sequences and related information.”

There are other fields-for example medical imaging/image analysis which might be considered part of bioinformatics. There is also a whole other discipline of biologically-inspired computation; genetic algorithms, AI, neural networks. Often these areas interact in strange ways. Neural networks, inspired by crude models of the functioning of nerve cells in the brain, are used in a program called PHD to predict, surprisingly accurately, the secondary structures of proteins from their primary sequences. What almost all bioinformatics has in common is the processing of large amounts of biologically-derived information, whether DNA sequences or breast X-rays.

Even though the three terms: bioinformatics, computational biology and bioinformation infrastructure are often times used interchangeably, broadly, the three may be defined as follows:

- **bioinformatics** refers to database-like activities, involving persistent sets of data that are maintained in a consistent state over essentially indefinite periods of time;

- **computational biology** encompasses the use of algorithmic tools to facilitate biological analyses; while

- **bioinformation infrastructure** comprises the entire collective of information management systems, analysis tools and communication networks supporting biology. Thus, the latter may be viewed as a computational scaffold of the former two.

National Center for Biotechnology Information (NCBI 2001) defines bioinformatics:

"Bioinformatics is the field of science in which biology, computer science, and information technology merges 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."

Applications of bioinformatics

- Sequence mapping of biomolecules (DNA, RNA, proteins).
- Identification of nucleotide sequences of functional genes.
- Finding of sites that can be cut by restriction enzymes.
- Designing of primer sequence for polymerase chain reaction.
- Prediction of functional gene products.
- To trace the evolutionary trees of genes.
- For the prediction of 3-dimensional structure of proteins.
- Molecular modelling of biomolecules.
- Designing of drugs for medical treatment.
- Handling of vast biological data which otherwise is not possible.
- Development of models for the functioning various cells, tissues and organs.

A Bioinformaticist versus a Bioinformatician (1999):

Bioinformatics has become a mainstay of genomics, proteomics, and all other *.omics (such as phenomics) that many information technology companies have entered the business or are considering entering the business, creating an IT (information technology) and BT (biotechnology) convergence.

Bioinformaticist	Bioinformatician
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.
A bioinformaticist is to *.omics as a mechanical engineer is to an automobile.	A bioinformatician is to *.omics as a technician is to an automobile

Some of the most important ones are listed below:

- Gene expression profiles
- Protein structure
- Protein interactions
- Microarrays (DNA chips)
- Functional analysis of biomolecules
- Drug designing.

Bioinformatics is largely (not exclusively) a computer-based discipline. Computers are in fact very essential to handle large volumes of biological data, their storage and retrieval. We have to accept the fact that there is no computer on earth (however advanced) which can store information, and perform the functions like a living cell. Thus a highly complex information technology lies right within the cells of an organism. This primarily includes the organism's genes and their dictates for the organism's biological processes and behaviour.

Broad Coverage of Bioinformatics:

Bioinformatics covers many specialized and advanced areas of biology.

- **Functional genomics:** Identification of genes and their respective functions.
- **Structural genomics:** Predictions related to functions of proteins.
- **Comparative genomics:** For understanding the genomes of different species of organisms.
- **DNA microarrays:** These are designed to measure the levels of gene expression in different tissues, various stages of development and in different diseases.
- **Medical informatics:** This involves the management of biomedical data with special referee to biomolecules, in vitro assays and clinical trials.

Bioinformatics and the Internet:

The internet is an international computer network. A computer network involves a group of computers that can communicate (usually over a telephone system) and exchange data between users. It is the internet protocol (IP) that determines how the packets of information are addressed and routed over the network. To access the internet, a computer must have the correct hardware (modem/ network card), appropriate software and permission for access to network. For this purpose, one has to subscribe to an internet service provider (ISP).

World Wide Web (www):

www involves the exchange of information over the internet using a programme called browser. The most widely used browsers are Internet explorer and Netscape navigator.

Components of Bioinformatics

1. Creation of databases:

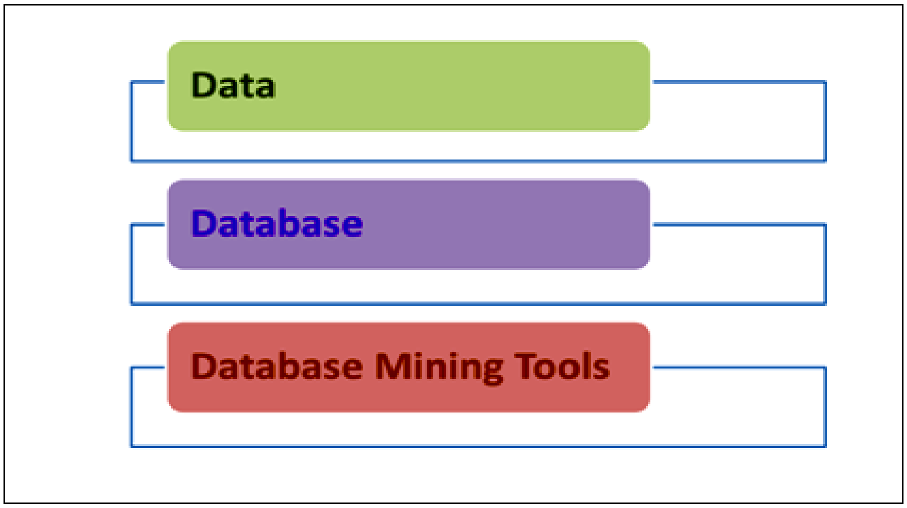
This involves the organizing, storage and management the biological data sets. The databases are accessible to researchers to know the existing information and submit new entries, e.g. protein sequence data bank for molecular structure. Databases will be of no use until analysed.

2. Development of algorithms and statistics:

This involves the development of tools and resources to determine the relationship among the members of large data sets e.g. comparison of protein sequence data with the already existing protein sequences.

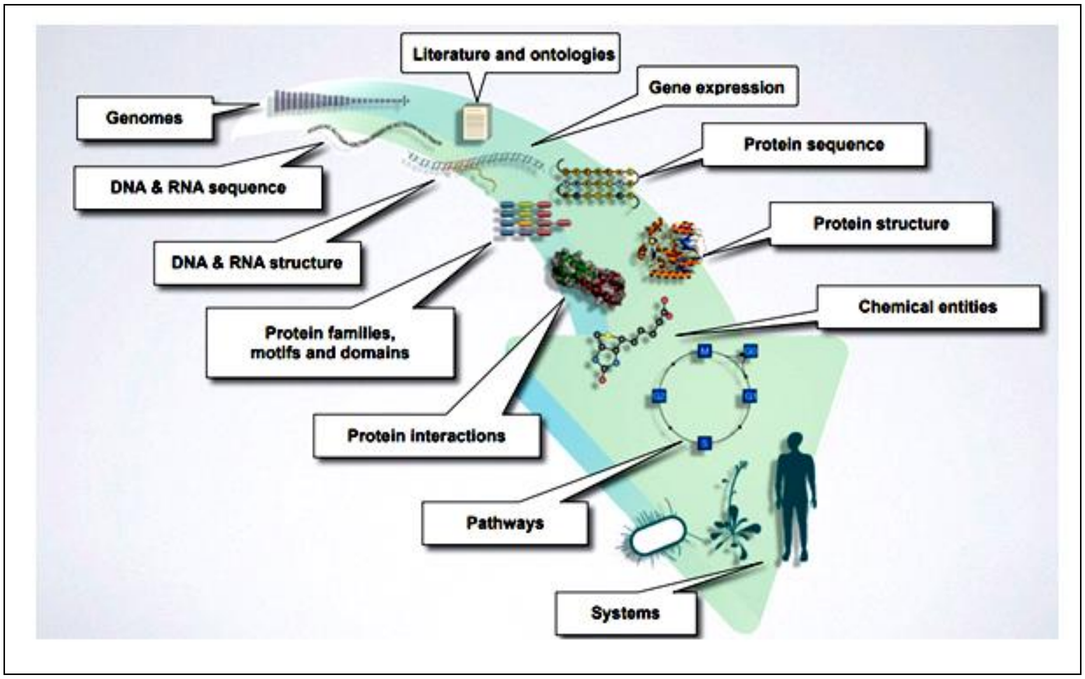
3. Analysis of data and interpretation:

The appropriate use of components 1 and 2 (given above) to analyse the data and interpret the results in a biologically meaningful manner. This includes DNA, RNA and protein sequences, protein structure, gene expression profiles and biochemical pathways.



Data

- ➢ Nucleic Acid Sequences
  - Raw DNA Sequences
  - Genomic sequence tags (GSTs)
  - cDNA sequences
  - Expressed sequence tags (ESTs)
  - Organellar DNA sequences
  - RNA Sequences
- ➢ Protein sequences
- ➢ Protein structures
- ➢ Metabolic pathways
- ➢ Gel pictures
- ➢ Literature



Databases

A database is a vast collection of data pertaining to a specific topic e.g. nucleotide sequence, protein sequence etc., in an electronic environment.

- ❖ They are heart of bioinformatics.
- ❖ Computerized storehouse of data (records).
- ❖ Allows extraction of specified records.
- ❖ Allows adding, changing, removing, and merging of records.
- ❖ Uses standardized formats.

Types of Database:

- ✓ Sequence Databases
- ✓ Structural Databases
- ✓ Enzyme Databases
- ✓ Micro-array Databases
- ✓ Clinical Database
- ✓ Pathway Databases
- ✓ Chemical Databases
- ✓ Integrated Databases
- ✓ Bibliographic Databases

Nucleotide Sequence Databases

❍ NCBI - GenBank: (www.ncbi.nlm.nih.gov/GenBank)
❍ EMBL: (www.ebi.ac.uk/embl)
❍ DDBJ: (www.ddbj.nig.ac.jp)

❖ The 3 databases are updated and exchanged on a daily basis and the accession numbers are consistent.

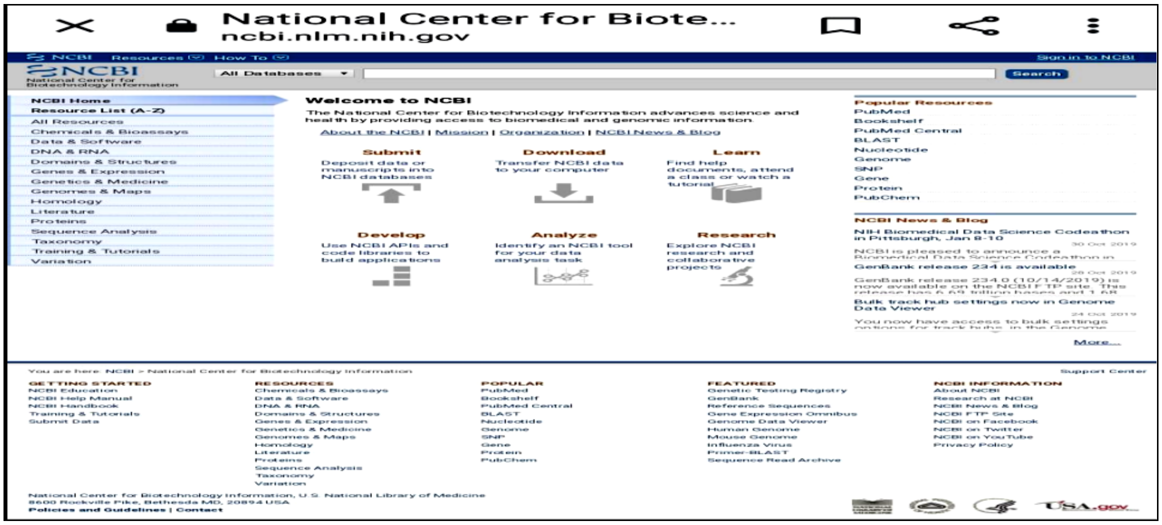
❖ There are no legal restriction in the usage of these databases. However, there are some patented sequences in the database.

❖ The International Nucleotide Sequence Database Collaboration (INSD)

National Center for Biotechnology Information (NCBI)

GenBank:

GenBank (Genetic Sequence Databank) is one of the fastest growing repositories of known genetic sequences. It has a flat file structure, that is an ASCII text file, readable by both humans and computers. In addition to sequence data, GenBank files contain information like accession numbers and gene names, phylogenetic classification and references to published literature.There are approximately 191,400,000 bases and 183,000 sequences as of June 1994.



National Center for Biote...
ncbi.nlm.nih.gov

NCBI Resources How To Sign in to NCBI

NCBI
National Center for Biotechnology Information

All Databases ▾ Search

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI News & Blog

NIH Biomedical Data Science Codeathon in Pittsburgh, Jan 8-10
30 Oct 2019
NCBI is pleased to announce a Biomedical Data Science Codeathon in

GenBank release 234 is available
28 Oct 2019
GenBank release 234.0 (10/14/2019) is now available on the NCBI FTP site. This release has 6.69 trillion bases and 1.68

Bulk track hub settings now in Genome Data Viewer
24 Oct 2019
You now have access to bulk settings options for track hubs in the Genome

More...

You are here: NCBI > National Center for Biotechnology Information
Support Center

GETTING STARTED
NCBI Education
NCBI Help Manual
NCBI Handbook
Training & Tutorials
Submit Data

RESOURCES
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Variation

POPULAR
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

FEATURED
Genetic Testing Registry
GenBank
Reference Sequences
Gene Expression Omnibus
Genome Data Viewer
Human Genome
Mouse Genome
Influenza Virus
Primer-BLAST
Sequence Read Archive

NCBI INFORMATION
About NCBI
Research at NCBI
NCBI News & Blog
NCBI FTP Site
NCBI on Facebook
NCBI on Twitter
NCBI on YouTube
Privacy Policy

National Center for Biotechnology Information, U.S. National Library of Medicine
8600 Rockville Pike, Bethesda MD, 20894 USA
Policies and Guidelines | Contact

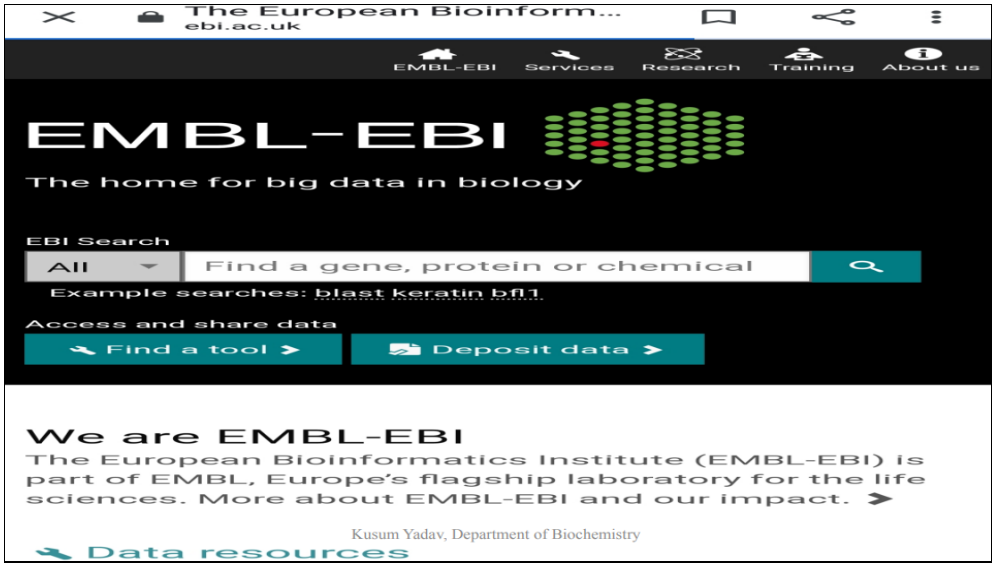
USA.gov

EMBL Database

The EMBL Nucleotide Sequence Database is a comprehensive database of DNA and RNA sequences collected from the scientific literature and patent applications and directly submitted from researchers and sequencing groups. Data collection is done in collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). The database currently doubles in size every 18 months and currently (June 1994) contains nearly 2 million bases from 182,615 sequence entries.

European Molecular Biology Laboratory (EMBL):

- ❖ Maintained by European Bioinformatics Institute (EBI)
  - ✓ GSS (genome survey sequences)
  - ✓ HTC (high-throughput c-DNA sequences)
  - ✓ HTG (high-throughput genomic sequences)
  - ✓ EST (expressed sequence tag)
  - ✓ Patents



EMBL-EBI

The home for big data in biology

Example searches: blast keratin bfl1

Access and share data

We are EMBL-EBI

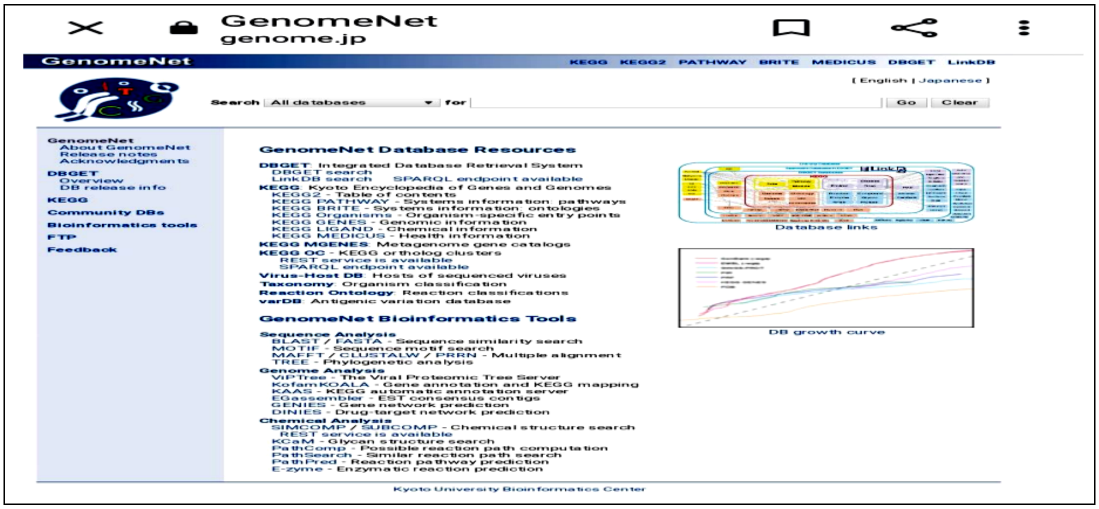
The European Bioinformatics Institute (EMBL-EBI) is part of EMBL, Europe's flagship laboratory for the life sciences. More about EMBL-EBI and our impact.

Kusum Yadav, Department of Biochemistry

Data resources

DDBJ (DNA Database of GenomNet, Japan)

- ✓ Developed in 1986 as a collaboration with EMBL and GenBank.
- ✓ Produced, maintained and distributed by the National Institute of Genetics, Japan.
- ✓ Sequences are submitted via Web based data submission tool.



GenomeNet Database Resources

DBGET: Integrated Database Retrieval System
DBGET search
LinkDB search SPARQL endpoint available
KEGG: Kyoto Encyclopedia of Genes and Genomes
KEGG2 - Table of contents
KEGG PATHWAY - Systems information: pathways
KEGG BRITE - Systems information: ontologies
KEGG Organisms - Organism-specific entry points
KEGG GENES - Genomic information
KEGG LIGAND - Chemical information
KEGG MEDICUS - Health information
KEGG MGENES: Metagenome gene catalogs
KEGG OC - KEGG ortholog clusters
REST service is available
SPARQL endpoint available
Virus-Host DB: Hosts of sequenced viruses
Taxonomy: Organism classification
Reaction Ontology: Reaction classifications
varDB: Antigenic variation database

GenomeNet Bioinformatics Tools

Sequence Analysis
BLAST / FASTA - Sequence similarity search
MOTIF - Sequence motif search
MAFFT / CLUSTALW / PRRN - Multiple alignment
TREE - Phylogenetic analysis
Genome Analysis
ViPTree - The Viral Proteomic Tree Server
KofamKOALA - Gene annotation and KEGG mapping
KAAS - KEGG automatic annotation server
EGassembler - EST consensus contigs
GENIES - Gene network prediction
DINIES - Drug-target network prediction
Chemical Analysis
SIMCOMP / SUBCOMP - Chemical structure search
REST service is available
KCaM - Glycan structure search
PathComp - Possible reaction path computation
PathSearch - Similar reaction path search
PathPred - Reaction pathway prediction
E-zyme - Enzymatic reaction prediction

Kyoto University Bioinformatics Center

Biological Database

A biological database is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system. A simple database might be a single file containing many records, each of which includes the same set of information. For example, a record associated with a nucleotide sequence database typically contains information such as contact name; the input sequence with a description of the type of molecule; the scientific name of the source organism from which it was isolated and often, literature citations associated with the sequence.

For researchers to benefit from the data stored in a database, two additional requirements must be met:

- Easy access to the information; and
- A method for extracting only that information needed to answer a specific biological question.

Primary Database

SwissProt:

This is a protein sequence database that provides a high level of integration with other databases and also has a very low level of redundancy (means less identical sequences are present in the database).

TrEMBL:

Computer annotated supplements of SWISS-PROT that contains all the translations of EMBL nucleotide entries not yet integrated in SWISS-PROT.

PIR-PSD:

PIR (Protein Information Resource) produces and distributes the PIR-International Protein Sequence Database (PSD). It is the most comprehensive and expertly annotated protein sequence database. The PIR serves the scientific community through on -line access, distributing magnetic tapes, and performing off- line sequence identification services for researchers. Release 40.00: March 31, 1994 - 67,423 entries - 19,747,297 residues.

Protein sequence databases are classified as primary, secondary and composite depending upon the content stored in them. PIR and SwissProt are primary databases that contain protein sequences as 'raw' data. Secondary databases (like Prosite) contain the information derived from protein sequences. Primary databases are combined and filtered to form non-redundant composite database

Secondary Database

Secondary database compile and filter sequence data from different primary database. These database contain information derived from protein sequences and help the user determine whether a new sequence belong to a known protein family**.**

PROSITE:

The PROSITE dictionary of sites and patterns in proteins prepared by Amos Bairoch at the University of Geneva.

PRINTS

PRINTS provides a compendium of protein fingerprints (groups of conserved motifs that characterise a protein family).

Now has a relational version, "PRINTS-S".

BLOCKS

BLOCK patterns without gaps in aligned protein families defined by PROSITE, found by pattern searching and statistical sampling algorithms.

Automatically determined un-gapped conserved segments.

Pfam

Database of protein families defined as domains.

For each domain, it contains a multiple alignment of a set of defining sequences and the other sequences in SWISS-PROT and TrEMBL that can be matched to the alignment**.**

EC-ENZYME:

The 'ENZYME' data bank contains the following data for each type of characterized enzyme for which an EC number has been provided: EC number, Recommended name, Alternative names, Catalytic activity, Cofactors, Pointers to the SWISS-PROT entrie(s) that correspond to the enzyme, Pointers to disease(s) associated with a deficiency of the enzyme.

PDB:

The X-ray crystallography Protein Data Bank (PDB), compiled at the Brookhaven National Laboratory.

GDB:

The GDB Human Genome Data Base supports biomedical research, clinical medicine, and professional and scientific education by providing for the storage and dissemination of data about genes and other DNA markers, map location, genetic disease and locus information, and bibliographic information.

OMIM:

The Mendelian Inheritance in Man data bank (MIM) is prepared by Victor Mc Kusick with the assistance of Claire A. Francomano and Stylianos E. Antonarakis at John Hopkins University.

Genethon Genome Databases

PHYSICAL MAP: computation of the human genetic map using DNA fragments in the form of YAC contigs. GENETIC MAP: production of micro- satellite probes and the localization of chromosomes, to create a genetic map to aid in the study of hereditary diseases. GENEXPRESS (cDNA): catalogue the transcripts required for protein synthesis obtained from specific tissues, for example neuromuscular tissues.

21 Bdb: LBL's Human Chr 21 database:

This is a W3 interface to LBL's ACeDB-style database for Chromosome 21, 21Bdb, using the ACeDB gateway software developed and provided by Guy Decoux at INRA.

MGD: The Mouse Genome Databases:

MGD is a comprehensive database of genetic information on the laboratory mouse. This initial release contains the following kinds of information: Loci (over 15,000 current and withdrawn symbols), Homologies (1300 mouse loci, 3500 loci from 40 mammalian species), Probes and Clones (about 10,000), PCR primers (currently 500 primer pairs), Bibliography (over 18,000 references), Experimental data (from 2400 published articles).

ACeDB (A Caenorhabditis elegans Database) :

Containing data from the Caenorhabditis Genetics Center (funded by the NIH National Center for Research Resources), the C. elegans genome project (funded by the MRC and NIH), and the worm community. Contacts: Mary O'Callaghan (moc@mrc-lmb.cam.ac.uk) and Richard Durbin.

ACeDB is also the name of the generic genome database software in use by an increasing number of genome projects. The software, as well as the C. elegans data, can be obtained via ftp.

ACeDB databases are available for the following species: C. elegans, Human Chromosome 21, Human Chromosome X, Drosophila melanogaster, mycobacteria, Arabidopsis, soybeans, rice, maize, grains, forest trees, Solanaceae, Aspergillus nidulans, Bos taurus, Gossypium hirsutum, Neurospora crassa, Saccharomyces cerevisiae, Schizosaccharomyces pombe, and Sorghum bicolor.

MEDLINE:

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. Citations include the English abstract when published with the article (approximately 70% of the current file).

The principal requirements on the public data services are:

- **Data quality** - data quality has to be of the highest priority. However, because the data services in most cases lack access to supporting data, the quality of the data must remain the primary responsibility of the submitter.

- **Supporting data** - database users will need to examine the primary experimental data, either in the database itself, or by following cross-references back to network-accessible laboratory databases.

- **Deep annotation** - deep, consistent annotation comprising supporting and ancillary information should be attached to each basic data object in the database.

- **Timeliness** - the basic data should be available on an Internet-accessible server within days (or hours) of publication or submission.

- **Integration** - each data object in the database should be cross-referenced to representation of the same or related biological entities in other databases. Data services should provide capabilities for following these links from one database or data service to another.

Data Mining

Data mining refers to extracting or mining knowledge from large amounts of data. In other words, Data mining is the science, art, and technology of discovering large and complex bodies of data in order to discover useful patterns. Theoreticians and practitioners are continually seeking improved techniques to make the process more efficient, cost-effective, and accurate. Many other terms carry a similar or slightly different meaning to data mining such as knowledge mining from data, knowledge extraction and data/pattern analysis data dredging.

Data mining treats as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. In others view data mining as simply an essential step in the process of knowledge discovery, in which intelligent methods are applied in order to extract data patterns.

Gregory Piatetsky-Shapiro coined the term “Knowledge Discovery in Databases” in 1989. However, the term ‘data mining’ became more popular in the business and press communities. Currently, Data Mining and Knowledge Discovery are used interchangeably.

Nowadays, data mining is used in almost all places where a large amount of data is stored and processed in this figure.

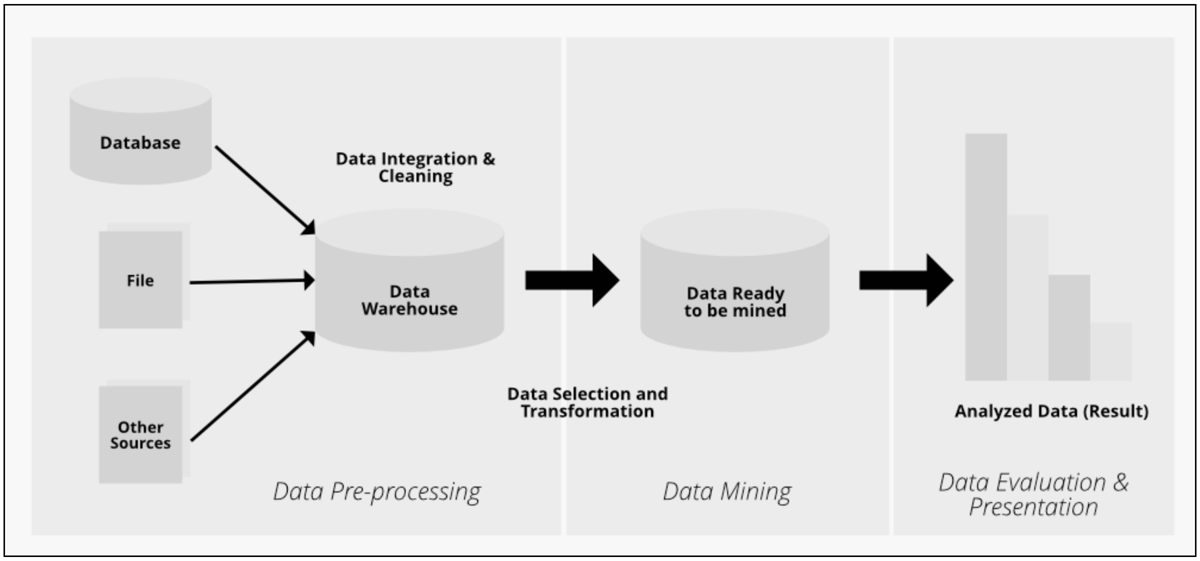


Figure 2: Data Mining Process

Knowledge Discovery From Data Consists of the Following Steps:

- Data cleaning (to remove noise or irrelevant data).
- Data integration (where multiple data sources may be combined).
- Data selection (where data relevant to the analysis task are retrieved from the database).
- Data transformation (where data are transmuted or consolidated into forms appropriate for mining by performing summary or aggregation functions, for sample).
- Data mining (an important process where intelligent methods are applied in order to extract data patterns).
- Pattern evaluation (to identify the fascinating patterns representing knowledge based on some interestingness measures).
- Knowledge presentation (where knowledge representation and visualization techniques are used to present the mined knowledge to the user).

Now we discuss here different types of Data Mining Techniques which are used to predict desire output.

Data Mining Techniques

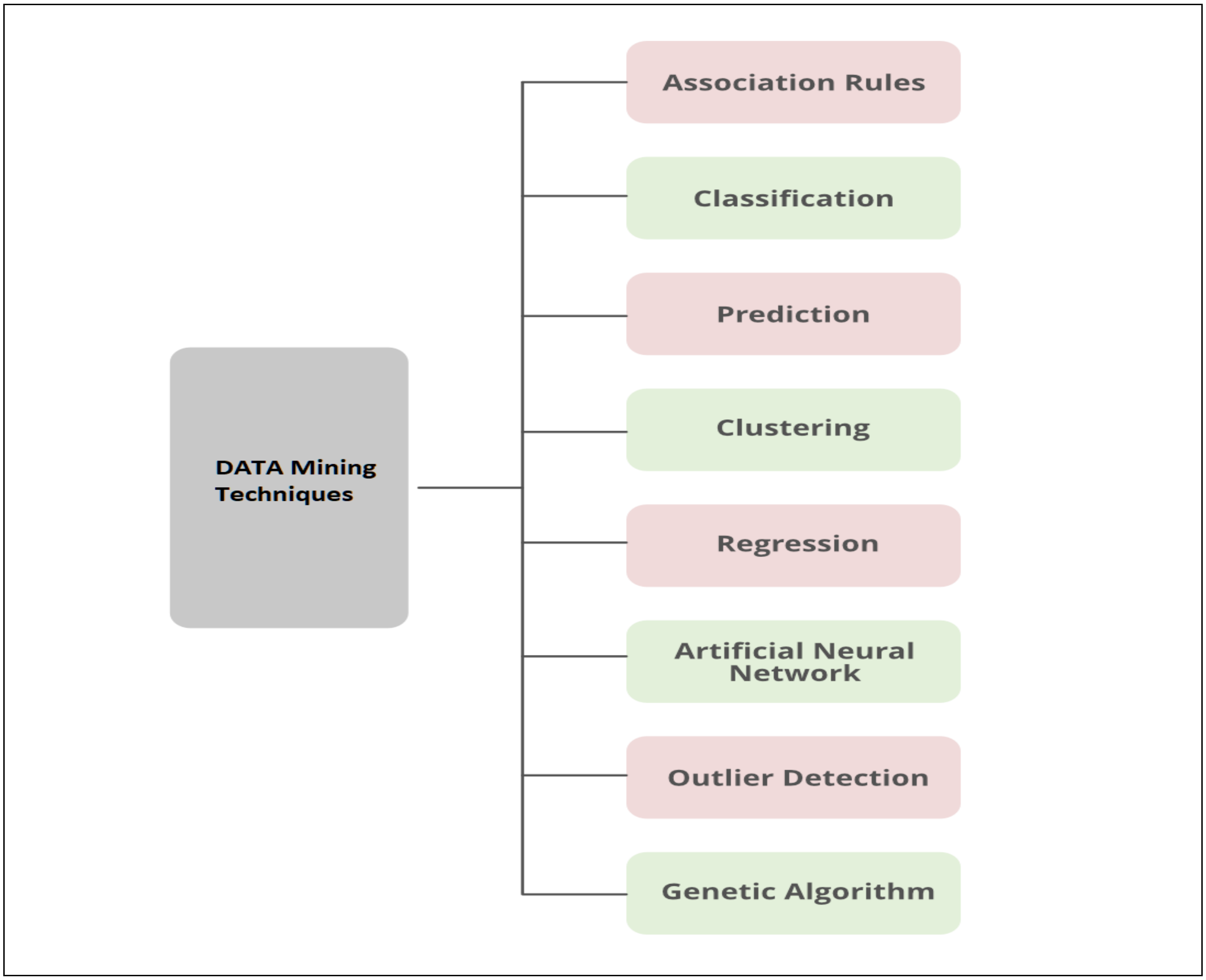


Figure 3: Data Mining Techniques

1. Association

Association analysis is the finding of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for a market basket or transaction data analysis. Association rule mining is a significant and exceptionally dynamic area of data mining research. One method of association-based classification, called associative classification, consists of two steps. In the main step, association instructions are generated using a modified version of the standard association rule mining algorithm known as Apriori. The second step constructs a classifier based on the association rules discovered.

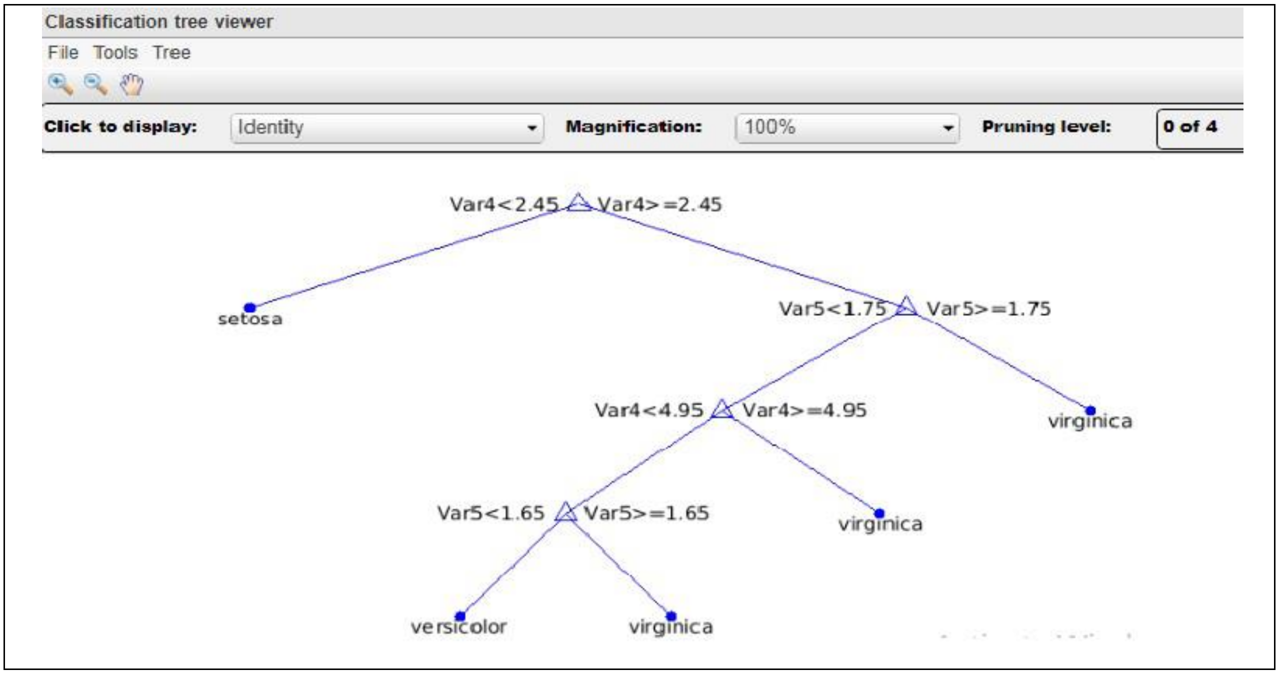
2. Classification

Classification is the processing of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The determined model depends on the investigation of a set of training data information (i.e. data objects whose class label is known). The derived model may be represented in various forms, such as classification (if – then) rules, decision trees, and neural networks. Data Mining has a different type of classifier:

- Decision Tree
- SVM(Support Vector Machine)
- Generalized Linear Models
- Bayesian classification
- Classification by Backpropagation
- K-NN Classifier
- Rule-Based Classification
- Frequent-Pattern Based Classification
- Rough set theory
- Fuzzy Logic

(i) **Decision Trees:**

A decision tree is a flow-chart-like tree structure, where each node represents a test on an attribute value, each branch denotes an outcome of a test, and tree leaves represent classes or class distributions. Decision trees can be easily transformed into classification rules. Decision tree enlistment is a nonparametric methodology for building classification models. In other words, it does not require any prior assumptions regarding the type of probability distribution satisfied by the class and other attributes. Decision trees, especially smaller size trees, are relatively easy to interpret. The accuracies of the trees are also comparable to two other classification techniques for a much simple data set. These provide an expressive representation for learning discrete-valued functions. However, they do not simplify well to certain types of Boolean problems.



Click to display: Identity

Magnification: 100%

Pruning level: 0 of 4

Figure 4: Decision Tree

This figure generated on the IRIS data set of the UCI machine repository. Basically, three different class labels available in the data set: Setosa, Versicolor, and Virginia.

(ii) **Support Vector Machine (SVM) Classifier Method:**

Support Vector Machines is a supervised learning strategy used for classification and additionally used for regression. When the output of the support vector machine is a continuous value, the learning methodology is claimed to perform regression; and once the learning methodology will predict a category label of the input object, it’s known as classification. The independent variables could or could not be quantitative. Kernel equations are functions that transform linearly non-separable information in one domain into another domain wherever the instances become linearly divisible. Kernel equations are also linear, quadratic, Gaussian, or anything that achieves this specific purpose. A linear classification technique may be a classifier that uses a linear function of its inputs to base its decision on. Applying the kernel equations arranges the information instances in such a way at intervals in the multi-dimensional space, that there is a hyper-plane that separates knowledge instances of one kind from those of another.

(iii) **Generalized Linear Models:**

Generalized Linear Models(GLM) is a statistical technique, for linear modeling.GLM provides extensive coefficient statistics and model statistics, as well as row diagnostics. It also supports confidence bounds.

(iv) **Bayesian Classification:**

Bayesian classifier is a statistical classifier. They can predict class membership probabilities, for instance, the probability that a given sample belongs to a particular class. Bayesian classification is created on the Bayes theorem. Studies comparing the classification algorithms have found a simple Bayesian classifier known as the naive Bayesian classifier to be comparable in performance with decision tree and neural network classifiers. Bayesian classifiers have also displayed high accuracy and speed when applied to large databases.

(v) **Classification By Back propagation:**

A Back propagation learns by iteratively processing a set of training samples, comparing the network's estimate for each sample with the actual known class label. For each training sample, weights are modified to minimize the mean squared error between the network's prediction and the actual class. These changes are made in the "backward" direction, i.e., from the output layer, through each concealed layer down to the first hidden layer (hence the name back propagation). Although it is not guaranteed, in general, the weights will finally converge, and the knowledge process stops.

(vi) **K-Nearest Neighbor (K-NN) Classifier Method:**

The k-nearest neighbor (K-NN) classifier is taken into account as an example-based classifier, which means that the training documents are used for comparison instead of an exact class illustration, like the class profiles utilized by other classifiers. As such, there's no real training section. once a new document has to be classified, the k most similar documents (neighbors) are found and if a large enough proportion of them are allotted to a precise class, the new document is also appointed to the present class, otherwise not. Additionally, finding the closest neighbors is quickened using traditional classification strategies.

(vii) **Rule-Based Classification:**

Rule-based classifications represent the knowledge in the form of If-Then rules. An assessment of a rule evaluated according to the accuracy and coverage of the classifier. If more than one rule is triggered then we need to conflict resolution in rule-based classification. Conflict resolution can be performed on three different parameters: Size ordering, Class-Based ordering, and rule-based ordering. There are some advantages of Rule-based classifier like:

- Rules are easier to understand than a large tree.
- Rules are mutually exclusive and exhaustive.
- Each attribute-value pair along a path forms conjunction: each leaf holds the class prediction.

(viii) **Frequent-Pattern Based Classification:**

Frequent pattern discovery (or FP discovery, FP mining, or Frequent itemset mining) is part of data mining. It describes the task of finding the most frequent and relevant patterns in large datasets. The idea was first presented for mining transaction databases. Frequent patterns are defined as subsets (item sets, subsequences, or substructures) that appear in a data set with a frequency no less than a user-specified or auto-determined threshold.

(ix) **Rough Set Theory:**

Rough set theory can be used for classification to discover structural relationships within imprecise or noisy data. It applies to discrete-valued features. Continuous-valued attributes must therefore be discrete prior to their use. Rough set theory is based on the establishment of equivalence classes within the given training data. All the data samples forming a similarity class are indiscernible, that is, the samples are equal with respect to the attributes describing the data. Rough sets can also be used for feature reduction (where attributes that do not contribute towards the classification of the given training data can be identified and removed), and relevance analysis (where the contribution or significance of each attribute is assessed with respect to the classification task). The problem of finding the minimal subsets (redacts) of attributes that can describe all the concepts in the given data set is NP-hard. However, algorithms to decrease the computation intensity have been proposed.

(x) **Fuzzy-Logic**:

Rule-based systems for classification have the disadvantage that they involve sharp cut-offs for continuous attributes. Fuzzy Logic is valuable for data mining frameworks performing grouping /classification. It provides the benefit of working at a high level of abstraction. In general, the usage of fuzzy logic in rule-based systems involves the following:

- Attribute values are changed to fuzzy values.
- For a given new data set /example, more than one fuzzy rule may apply. Every applicable rule contributes a vote for membership in the categories. Typically, the truth values for each projected category are summed.

3. Prediction

Data Prediction is a two-step process, similar to that of data classification. Although, for prediction, we do not utilize the phrasing of “Class label attribute” because the attribute for which values are being predicted is consistently valued(ordered) instead of categorical (discrete-esteemed and unordered). The attribute can be referred to simply as the predicted attribute. Prediction can be viewed as the construction and use of a model to assess the class of an unlabeled object, or to assess the value or value ranges of an attribute that a given object is likely to have.

4. Clustering

Unlike classification and prediction, which analyze class-labeled data objects or attributes, clustering analyzes data objects without consulting an identified class label. In general, the class labels do not exist in the training data simply because they are not known to begin with. Clustering can be used to generate these labels. The objects are clustered based on the principle of maximizing the intra-class similarity and minimizing the interclass similarity. That is, clusters of objects are created so that objects inside a cluster have high similarity in contrast with each other, but are different objects in other clusters. Each Cluster that is generated can be seen as a class of objects, from which rules can be inferred. Clustering can also facilitate classification formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

5. Regression

Regression can be defined as a statistical modeling method in which previously obtained data is used to predicting a continuous quantity for new observations. This classifier is also known as the Continuous Value Classifier. There are two types of regression models: Linear regression and multiple linear regression models.

6. Artificial Neural network (ANN) Classifier Method

An artificial neural network (ANN) also referred to as simply a "Neural Network" (NN), could be a process model supported by biological neural networks. It consists of an interconnected collection of artificial neurons. A neural network is a set of connected input/output units where each connection has a weight associated with it. During the knowledge phase, the network acquires by adjusting the weights to be able to predict the correct class label of the input samples. Neural network learning is also denoted as connectionist learning due to the connections between units. Neural networks involve long training times and are therefore more appropriate for applications where this is feasible. They require a number of parameters that are typically best determined empirically, such as the network topology or "structure". Neural networks have been criticized for their poor interpretability since it is difficult for humans to take the symbolic meaning behind the learned weights. These features firstly made neural networks less desirable for data mining.

An artificial neural network is an adjective system that changes its structure-supported information that flows through the artificial network during a learning section. The ANN relies on the principle of learning by example. There are two classical types of neural networks, perceptron and also multilayer perceptron.

7. Outlier Detection

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are Outliers. The investigation of OUTLIER data is known as OUTLIER MINING. An outlier may be detected using statistical tests which assume a distribution or probability model for the data, or using distance measures where objects having a small fraction of “close” neighbors in space are considered outliers.

8. Genetic Algorithm

Genetic algorithms are adaptive heuristic search algorithms that belong to the larger part of evolutionary algorithms. Genetic algorithms are based on the ideas of natural selection and genetics. These are intelligent exploitation of random search provided with historical data to direct the search into the region of better performance in solution space. They are commonly used to generate high-quality solutions for optimization problems and search problems. Genetic algorithms simulate the process of natural selection which means those species who can adapt to changes in their environment are able to survive and reproduce and go to the next generation. In simple words, they simulate “survival of the fittest” among individuals of consecutive generations for solving a problem. Each generation consist of a population of individuals and each individual represents a point in search space and possible solution.

Bioinformatics Tools

Following are the some of the important tools for bioinformatics:

Bioinformatics Research Area	Tool (Application)	References
Sequence alignment	BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi
	CS- BLAST	ftp://toolkit.lmb.uni
	HMMER	http://hmmer.janelia.org/
	FASTA	www.ebi.ac.uk/fasta33
Multiple sequence alignment	MSAProbs	http://msaprobs.sourceforge.net/
	DNA Alignment	http://www.fluxus-engineering.com/align.htm
	MultAlin	http://multalin.toulouse.inra.fr/multalin/multalin.html
	DiAlign	http://bibiserv.techfak.uni-bielefeld.de/dialign/
Gene Finding	GenScan	genes.mit.edu/GENSCAN.html
	GenomeScan	http://genes.mit.edu/genomescan.html
	GeneMark	http://exon.biology.gatech.edu/
Protein Domain Analysis	Pfam	http://pfam.sanger.ac.uk/
	BLOCKS	http://blocks.fhcrc.org/
	ProDom	http://prodom.prabi.fr/prodom/current/html/home.php
Pattern Identification	Gibbs Sampler	http://bayesweb.wadsworth.org/gibbs/gibbs.html
	AlignACE	http://atlas.med.harvard.edu/
	MEME	http://meme.sdsc.edu/
Genomic Analysis	SLAM	http://bio.math.berkeley.edu/slam/
	Multiz	http://www.bx.psu.edu/miller_lab/
Motif finding	MEME/MAST	http://meme.sdsc.edu
	eMOTIF	http://motif.stanford.edu

Need for Data Mining in Bioinformatics

The entire human genome, the complete set of genetic information within each human cell has now been determined. Understanding these genetic instructions promises to allow scientists to better understand the nature of diseases and their cures, to identify the mechanisms underlying biological processes such as growth and ageing and to clearly track our evolution and its relationship with other species. The key obstacle lying between investigators and the knowledge they seek is the sheer volume of data available. This is evident from the following figure which shows the rapid increase in the number of base pairs and DNA sequences in the repository of GenBank.

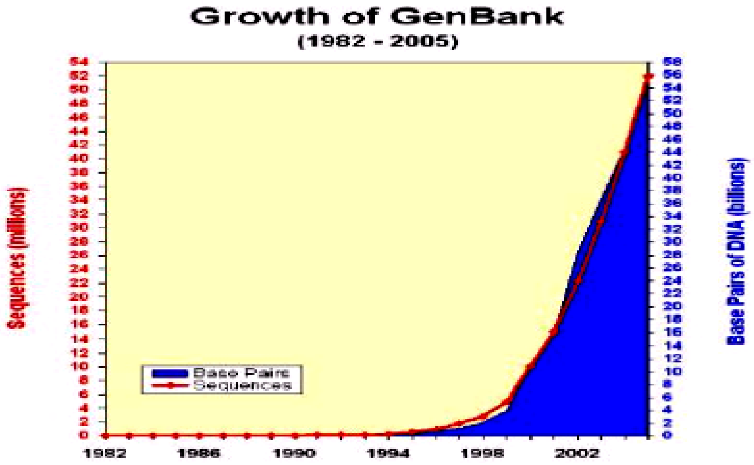


Figure 5: Growth of Genbank

Biologists, like most natural scientists, are trained primarily to gather new information. Until recently, biology lacked the tools to analyze massive repositories of information such as the human genome database. Luckily, the discipline of computer science has been developing methods and approaches well suited to help biologists manage and analyze the incredible amounts of data that promise to profoundly improve the human condition. Data mining is one such technology.

Biological Data Analysis

Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis −

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

Applications of Data Mining:

Applications of data mining to bioinformatics include

- Gene finding
- protein function domain detection
- Function motif detection
- Protein function inference
- Disease diagnosis
- Disease prognosis
- Disease treatment optimization
- Protein and gene interaction network
- Reconstruction
- Data cleansing
- Protein sub-cellular location prediction
- Analysis of protein and DNA sequences

For example, microarray technologies are used to predict a patient’s outcome. On the basis of patients’ genotypic microarray data, their survival time and risk of tumor metastasis or recurrence can be estimated. Machine learning can be used for peptide identification through mass spectroscopy. Correlation among fragment ions in a tandem mass spectrum is crucial in reducing stochastic mismatches for peptide identification by database searching. An efficient scoring algorithm that considers the correlative information in a tunable and comprehensive manner is highly desirable.