Bioinformatics

- **Proteomic : Genomics :-**
  Genomics is the study of the entire set of genes in the genome of a cell where as proteomics is the study of the entire set of proteins produced by cell

following biological analysis are performed by data mining :-

- Protein structure prediction
- Gene classification
- Analysis of mutations in cancer
- Gene Expressions

Data Mining :- It is the method extracting information for the use of learning patterns and models from large extensive datasets.

Data mining involves

(i) Machine learning
(ii) Statistics
(iii) Artificial intelligence.
(iv) database sets
(v) Pattern recognition and visualisation

KDD (Knowledge discovery in databases)
IDA (Intelligent Data analysis)

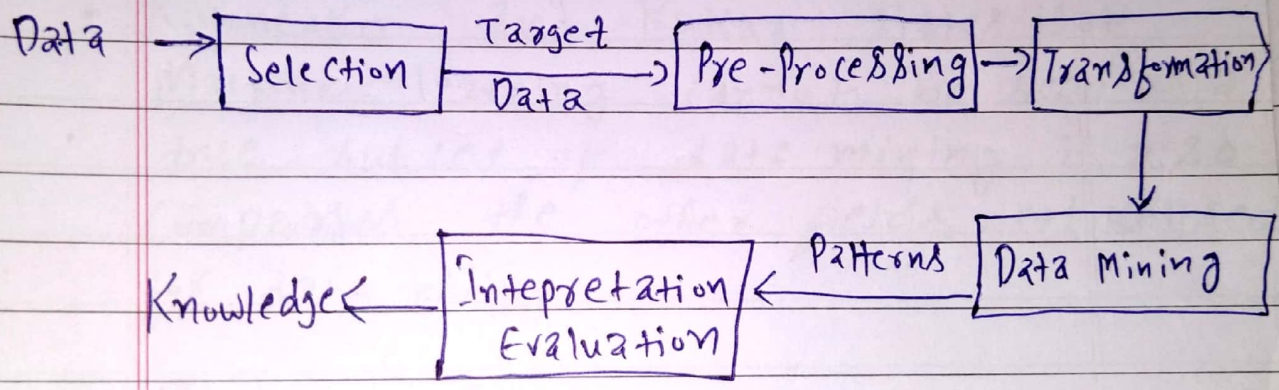
Application of Data mining and Machine learning models :-

- Kononenko and Kukar identify

⇒ Machine learning systems make rules, functions, relations, equation systems, probability distributions and other knowledge representations.

Aim of Data mining :- forecasting, Validation diagnosis and Simulations.

- process of Knowledge Discovery through Data mining :-



There is currently no standard framework of carrying out data mining.

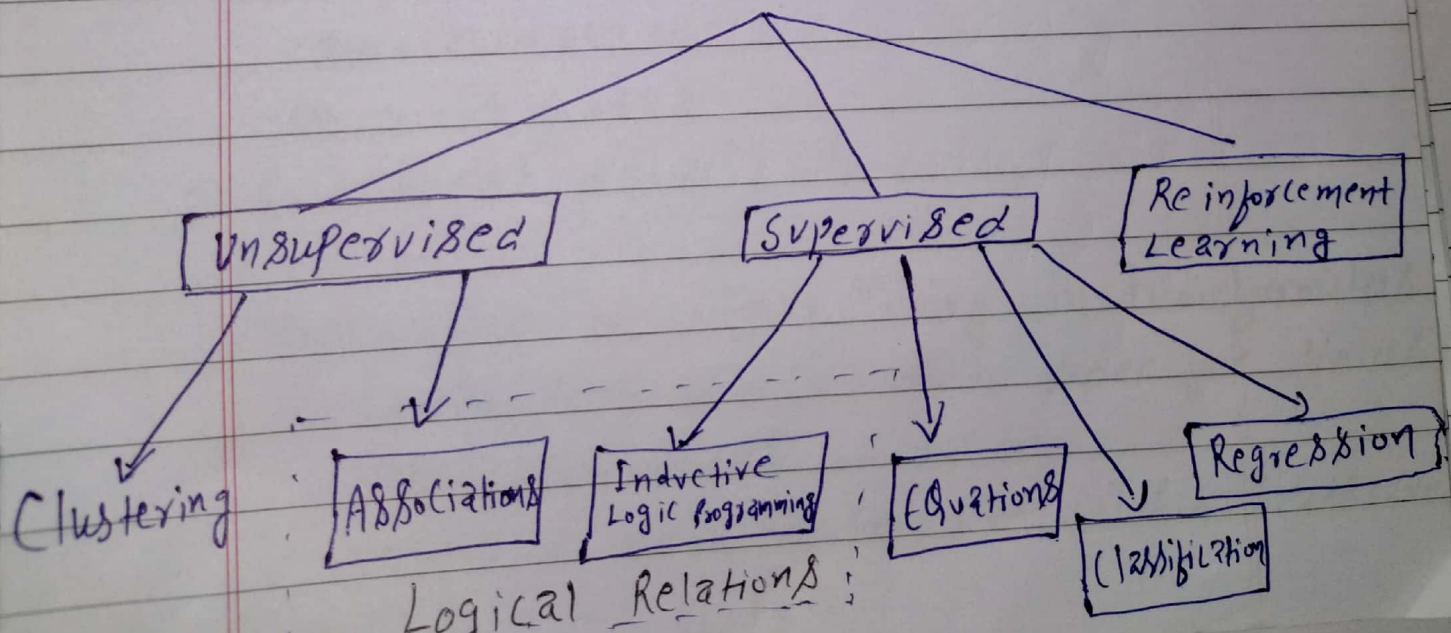
- CRISP-DM (Cross Industry Standard Process for Data mining)
It defines one standard framework for the process of data mining across multiple industries containing phases, generic tasks, specialised tasks and process instances.

Main tasks for data mining are.

1. Classification :- classifies a data item to a predefined class
2. Estimation :- Determining a value for unknown continuous variables
3. Prediction :- Records classified according to estimated future behaviour.
4. Association :- Defining items that are together
5. Clustering :- Defining a population into subgroups or clusters.
6. Description and visualisation :- Representing data

- Kononenko and Kukar states that Machine learning cannot be seen as a true subset of data mining it also compasses the other fields, not utilised for data mining.

Methods of Machine Learning



Bioinformatics:- It deals with the storage, gathering, simulation and analysis of biological data for the use of informatic tools such as data mining.

Fogel, Corne and Pan (2008) define as:- Research, development or application of Computational tools and approaches for expanding the use of biological, medical, behavioural or health data, including those to acquire, store, organise, archive analyse or visulise the data

Tramontano (2007) defines
We could define bioinformatics as the science that analyzes biological data with Computer tools in order to formulate hypotheses on the processes underlying life

Bioinformatic data is divided into three categories:-
Sequence data:-
Structural data:-
functional data.

Application of Data Mining in bioinformatics
- proactive research within specific fields of the biomedical industry.
- To discover new treatment within heathcare and knowledge of life

Razza(2010) propose following application
Gene finding
protein function domain detection
function motify detection
protein function inference

- To predict sequence outputs and create a hypothesis based on results.

Motif	Domain
• Super Secondary Structure	Tertiary Structure
• formed by the Connected alpha helices and beta Sheets through loops.	formed by formation of disulfide bridges, ionic bonds, H-bonds between amino acid side chains
• Are not Stable independtly	Stable Independently
• Similar functions as protein family	Unique functions
• Mainly have a structural function in the protein Structure	mainly have functional importance

Upregulation and Downregulation of enzyme or gene.

=> Upregulation occurs when a cell produces more receptors, the cell decreases its degradation of receptors or by activating already present receptors

Down regulation is when a cell decreases its sensitivity to a hormone by decreasing the amount of available receptors.

Gene balance hypothesis states that the stoichiometry of members of multi-subunit complexes affects the function of the whole due to the kinetics and mode of assembly.

Hormone clearance is the process of lowering hormone levels in the blood through two mechanism:- decrease secretion of a hormone or increases the degradation of a hormone.

Application of data mining in bioinformatics

1. Sequence analysis:-

This operation consists of finding which part of the biological sequences are alike and which part differs during medical analysis and genome mapping process.

The sequence analysis implies subjecting a DNA or Peptide sequence to sequence alignment, sequence databases, repeated sequence searches, or other bioinformatic methods on a computer.

2. Genome annotation

It is the process of marking the genes and other biological features in a DNA sequence. The first genome annotation system was designed by Dr. Owen White in 1995.

3. Analysis of gene expression

The expressions of many genes can be determined by measuring mRNA levels with various techniques such as microarrays expressed sequence tag cDNA (EST) sequencing, Serial analysis of gene expression (SAGE) tag sequencing, massively Parallel Signature Sequencing (MPSS), or various applications of multiplexed in-situ hybridization etc.

4. Analysis of Protein Expression

Protein expression is one of the best clues of actual gene activity since proteins are usually final catalysts of cell activity. Protein microarrays and High throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample.

5. Analysis of Mutations in Cancer:- oligonucleotide microarrays to identify chromosomal gains and losses and single nucleotide polymorphisms array to detect known point mutations.

6. Protein Structure Prediction:- The amino acid sequence of a protein can be easily determined from the sequence on the gene that codes for it. Knowledge of this structure is vital in understanding of Protein function.

7. Comparative genomics:- It is study of the relationship of genome structure and function across different biological species. Gene finding is discovery of new, non-coding functional elements of the genome,
Comparative genomics exploits both similarities and differences in the proteins, RNA and regulatory regions of different organisms.

8. Modelling biological Systems:- It is a significant task of systems biology and mathematical biology. Computational systems biology aims to develop and use efficient algorithms, data structures, visualization

and Communication tools for the integration of large quantities of biological data with the goal of Computer modelling. It involves the use of computer simulations of biological systems, like Cellular subsystems such as the network of metabolites and enzymes, signal transduction pathways and gene regulatory networks to both analyze and visualize the complex connections of these cellular processes.

9. High-throughput image analysis
Computational technologies are used to accelerate or fully automate the processing, qualification and analysis of large-amounts of high-information content biomedical images.

10. Protein protein docking:-
It is practical to predict possible protein protein interactions only based on the 3D shapes, without doing protein-protein interaction experiments.

7. motif finding:- MEME/MAST, eMOTIF.

Bioinformatics tools

1. Sequence alignment:- Blast, CM-blast, HMMER, FASTA
2. Multiple " " :- MSA Probs, MultiAlin, DiAlin, DNA
3. Gene finding :- GeneScan, genecome scan, GeneMark.
4. Protein domain analy:- Pfam BLOCKS, ProDom
5. Pattern identification:- Gibbs sampler, Align ACE, MEME.
6. Genome analysis :- SIAM, MUSTI2

BIG Data In bioinformatics:

Big data describes a large volume of data in bioinformatics and Computational biology. It represents a new paradigm that transforms the studies to a large-scale research.

The role of big data in bioinformatics
To provide repositories of data.
Better Computing facilities
Data manipulation tools to analyze data

Parallel Computing allows executing algorithms simultaneously on a cluster of machines or supercomputers.

Google -> MapReduce novel Parallel Computing model
Apache :- Hadoop (open source MapReduce package) : It also provides cloud Computing facilities for centralized data storage and provides remote access to them

Big Data technologies or tool categorized into four :-

1. Data Storage and Retrieval:-

(i) The sequencing data obtained has need to be mapped to specify reference genomes to further analysis. for this purpose, cloud burst a parallel computing model is used.

(ii) Contrain :- for assembling large genomes

(iii) Crossbow :- for identifying SNPs from sequence datasets

(iv) DistMap :- A toolkit for distributed short read mapping on a Hadoop Cluster

(v) Seqware: (to access large-scale whole genome datasets

(vi) Read Annotation Pipeline :- (by DDBJ, Cloud-based pipeline to analyse NGS data)

(vii) Hydra (for processing large peptide and spectra databases)

2. Error Identification.

(i) SAMQA :- which identifies errors and ensures that large-scale genomic data meet the minimum Quality standard

(ii) ART :- which simulates data for three major sequencing platforms viz: Sequen Illumina and SoliD.

(iii) CLoudRS

3. Data analysis
(i) GATK (Genome analysis Toolkit)
(ii) Array Express Archive of functional Genomics data repository
(iii) BlueSNP : It is used for genome-wide association

4. Platform Integration Deployment
(i) Seqpig :- It reduces the technological skill required to use MapReduce by reading large formatted files to feed analysis applications.
(ii) CloVR :- It is a sequencing analysis package distributed through a virtual machine
(iii) CloudBio Linux.

NGS (Next Generation Sequencing)

→ fragmentation
→ Ligation of DNA adaptors
→ Denaturation
→ Binding to chip
→ DNA Amplification by PCR (Denature)
→ Washing
→ Bridge Amplification
→ Denaturation
Repeated these steps.
Reverse strand is cleaved
Sequencing by synthesis
(RT-PCR)

BIAST

- Blast can do Multiple sequence alignmen
- Nucleotide blast
  Nucleotide Query ⇒ Nucleotide database

Basic Local Alignment Search tool
Blast finds regions of similarity betwee biological sequences. The program compare nucleotide or protein sequence to sequenc databases and calculates the statical significance.

Nucleotide blast
Nucleotide Query → Nucleotide database.
Blast X ⇒ translated nucleotide Query → protein database.
tblastn ⇒ protein query → Translated nuclotide database
protein blast ⇒ protein Query → protein database.

Specialized blast
Smart blast :- find protein highly similar to your query.

Primer blast :- Design primer specific to your PCR template

PCR test:-

RNA extraction

⇓

Reverse transcription

↓

Primer DNA polymerase Probe

↓

RT-PCR amplification.

↓

Results.

Global align :-

Compare two sequence across their entire span (needleman wunsch)

Pairwise sequence alignment:-

It is used to identify regions of similarity that may indicate functional, structural and/or evolutionary relationships between two biological sequence (protein or Nucleic acid)

MSA (Multiple sequence alignment) is the alignment of three or more biological sequence of similar length. From the output of MSA homology and evolutionary relationships studied.

Global alignment.
end to end alignment.
needle (EMBOSS)
It creates an optimal global Alor of two sequences using needleman wunsch algorithm.

Streteher.
It was modification of needleman-Wunsch algorithm that allows larger sequences to be globally aligned.

Local Alignment
Local alignment tools find one or more alignments describing the most similar regions within the sequences to be aligned. They can align protein and nucleotide sequences.

✓ Water (EMBOSS)
It was the Smith-Waterman algorithm (modified for speed enhancements) to calculate the local alignment of two sequences.

✓ Matcher (EMBOSS)
It identify local similarities between two sequences using a rigorous algorithm based on LALIGN Application.

✓ LALIGN
It finds internal duplications by calculating non-intersecting local alignments of protein or DNA sequences.

Genomic alignment:-

It finds Concentrate on DNA alignments while accounting for characteristics present in genomic data.

- Genewise :- It Compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors.
- CD-Search :- find Conserved domains in your sequence.
- Ig Blast :- Search immunoglobulins and T Cell receptor sequences.
- VecScreen :- Search sequences for vector Contamination.
- CDART :- find sequene with similar Conserved domain architecture.
- Mole-Blast -> Establish taxonomy for unCultured or environmental sequences

✓ Dust
A programme for filtering low complexity regions from nucleic acid sequences.

✓ E-value (Expectation or Expect value)
It represents the number of different alignment with scores equivalent to or better than S that is expected to occur in a database search by chance.

lower E value :- more significant score and alignment

filtering :- It also called masking, removes regions of sequence having characteristics that may lead to spurious high scores.

$$S = \sum (\text{identities, mismatches}) - \sum (\text{gap Penalties})$$

$$\text{Score} = \text{MAX}(S)$$

=) Expression Atlas

=) Pubmed :-

Rasbin - Rasmol for windows

Pdb → protein

open Rasmol - 2 Software.

It is a programme for molecular graphics visualisation originally developed by Roger Sayle

It is freely available.

NCBI - Sequence database (nucleotid)
PDB - structural database (Protein)

Pepsin

- What is slab mode?

Gromacs :- Molecular dynamic of any molecule particularly Protein Molecule.

Simulate the study