EDITION 2020

RITESH JAISWAL
ASTHA DUBEY

BASICS OF BIOINFORMATICS-II

BASED ON AKTU CURRICULUM

STUDY MATERIAL
FOR EVEN SEMESTER

BASICS OF BIOINFORMATICS-II

Edition-2020

Edited by

Mr. Ritesh Jaiswal

M.tech (Bioinformatics) SHUATS, Allahabad

Ms. Astha Dubey

B.tech (biotech) AITM ,Varanasi

preface

This study material will be helpful for all the students who are enrolled for the B.tech (biotechnology) programs who are in even semester and are going to give their exams soon. This study material is totally based on AKTU CURRICULUM and I and my student have tried to cover all the topics of the syllabus and notes is also arranged for the students in simple language to promote their understanding also. I am very thankful to my student who is pursuing her B.tech has helped me allot.

Ritesh Jaiswal

CONTENT

Unit I
• Inference problems and techniques for molecular biology.
• Homology identification,
• Genomic sequence annotation
• ORFs identification
• Biological network identification
• Microarray data analysis
• Protein function prediction
• Next generation sequencing
• Protein structure prediction (Secondary and Tertiary structure prediction)
Unit II
• Basics of RNA
• Features of RNA Secondary Structure
• RNA structure prediction methods:
a. Based on self-complementary regions in RNA sequence
b. Minimum free energy methods
• Suboptimal structure prediction by MFOLD
• Prediction based on finding most probable structure and Sequence co-variance method.
• Application of RNA structure modeling.
Unit III
• Machine learning
• Decision tree induction
• Artificial Neural Networks
• Hidden Markov Models
• Genetic Algorithms
• Simulated Annealing
• Support vector machines
• The relation between statistics and machine learning
• Evaluation of prediction methods: Parametric and Non-parametric tests
• Cross-validation and empirical significance testing (empirical cycle)
• Clustering (Hierarchical and K-mean).

Unit IV
• Basic concept of Force field in molecular modeling (Potential energy calculation)
• Overview of key computational simulation techniques
• Introduction to simulation
• Computer simulation techniques
• Types of computer simulation (Continuous, Discrete-event and Hybrid simulation)
• Differential equation solvers,
• Parameter estimation, and Sensitivity analysis
Unit V
• Overview of key techniques for the management of large document collections and the biological literature
• Document clustering
• Information retrieval system
• Natural Language Processing: Introduction
• Major areas of NLP
• Natural language information extraction
• Insilico Drug Designing
• Major steps in Drug Designing
• Ligand and Structure based drug designing
• Protein-ligand docking
• QSAR Modeling
• Pharmacodynamics (Efficacy & Potency)
• Pharmacokinetics (ADME)
• Lipinski's rule of five
• Pharmacogenomics.

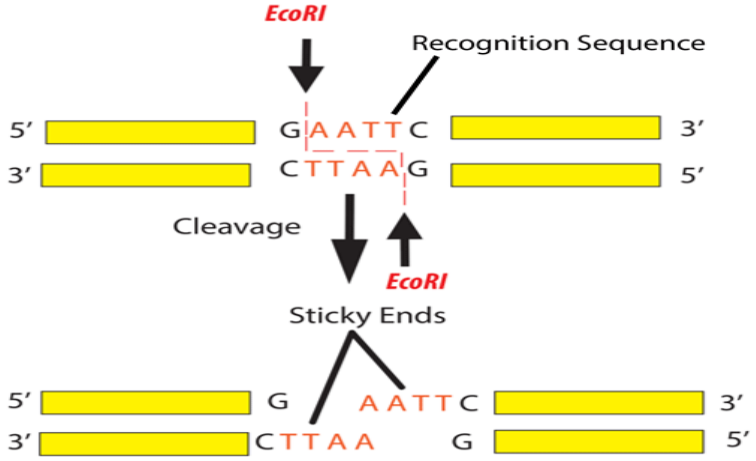
UNIT –I

Inferences problems and techniques in molecular biology-

1. **Enzymes in molecular biology**

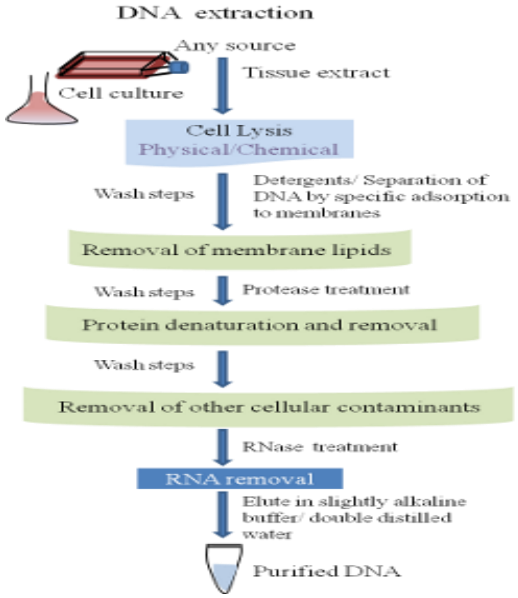
   Molecular biology is a study of different enzymes which are responsible for various functions.

   Ex – type II restriction enzyme plays a very important role in molecular biology. It cuts the DNA at recognized sequence site and divide it into fragments.

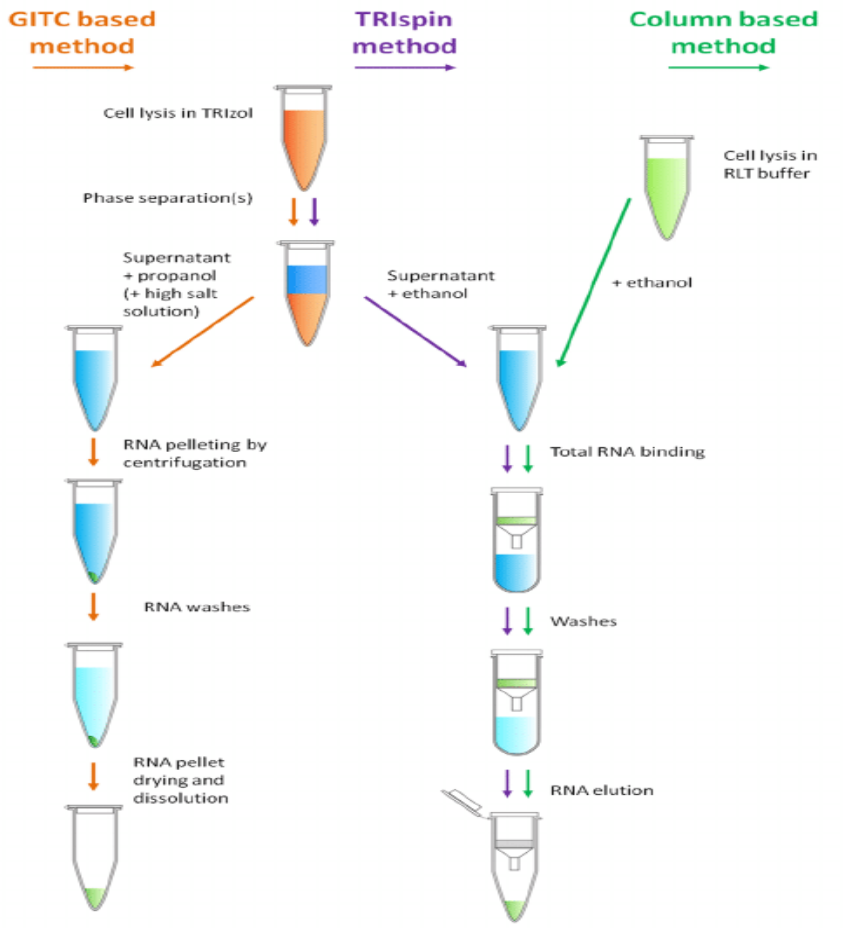


2. **Isolation and separation of nucleic acids-**

   a) Isolation of DNA-

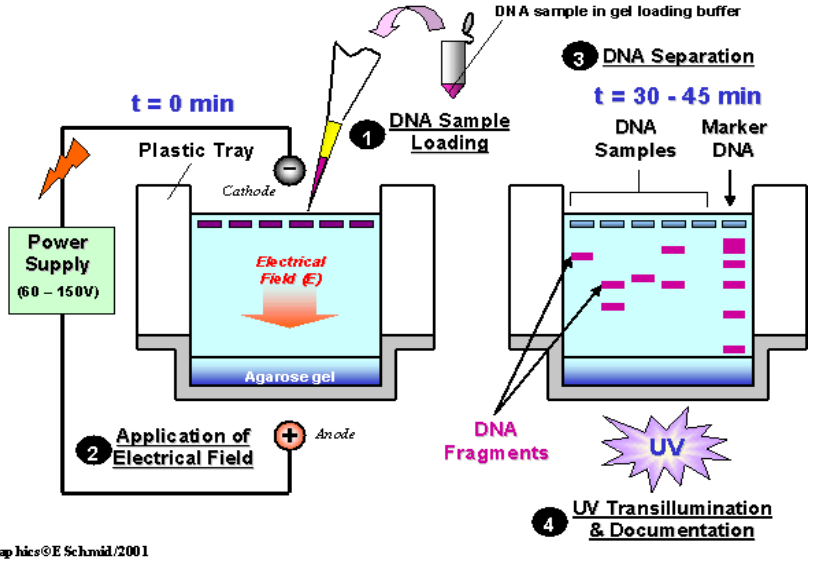


b). Isolation of RNA-



c).Gel electrophoresis is a method for separation and analysis of macromolecules (DNA, RNA and proteins) and their fragments, based on their size and charge. It is used in clinical chemistry to separate proteins by charge or size (IEF agarose, essentially size independent) and in biochemistry and molecular biology to separate a mixed population of DNA and RNA fragments by length, to estimate the size of DNA and RNA fragments or to separate proteins by charge.

Nucleic acid molecules are separated by applying an electric field to move the negatively charged molecules through a matrix of agarose or other substances. Shorter molecules move faster and migrate farther than longer ones because shorter molecules migrate more easily through the pores of the gel. This phenomenon is called sieving.

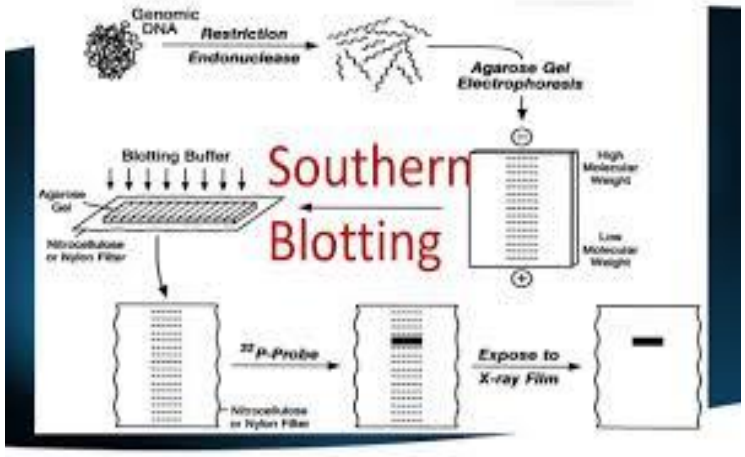


3. Blotting techniques-:

A **blot**, in molecular biology and genetics, is a method of transferring proteins, DNA or RNA onto a carrier (for example, a nitrocellulose, polyvinylidene fluoride or nylon membrane). In many instances, this is done after a gel electrophoresis, transferring the molecules from the gel onto the blotting membrane, and other times adding the samples directly onto the membrane. After the blotting, the transferred proteins, DNA or RNA are then visualized by colorant staining (for example, silver staining proteins) autoradiographic visualization of Radiolabelled molecules (performed before the blot), or specific labelling of some proteins or nucleic acids. The latter is done with antibodies or hybridization probes that bind only to some molecules of the blot and have an enzyme joined to them. After proper washing, this enzymatic activity (and so, the molecules we search in the blot) is visualized by incubation with proper reactive, rendering either a colored deposit on the blot or a chemiluminescent reaction which is registered by photographic film.

	Southern Blot	Northern Blot	Western Blot
Target molecule	DNA	RNA	Protein
Sample preparation	DNA extraction enzymatic digestion	RNA isolation	Protein extraction
Separation	Electrophoresis	Electrophoresis	Electrophoresis
Membrane material	Nylon	Nylon	Nitrocellulose or PVDF
Probe	Nucleic acid probe with sequence homologous to target	RNA, DNA, or oligodeoxynucleotide	Primary antibody
Probe label	Radiolabel, enzyme	Radiolabel, enzyme	Enzyme
Detection methods	X-ray film, chemiluminescence	X-ray film, chemiluminescence	Film, cooled CCD, camera, LED, or infrared imaging system

Table 1: Comparing Southern, Northern, and Western Blots.



Homology modeling-:

-Homology modeling, also known as **comparative modeling** of protein, refers to constructing an atomic-resolution model of the "target" protein from its amino acid sequence and an experimental three-dimensional structure of a related homologous protein (the "template").

-Homology modeling relies on the identification of one or more known protein structures likely to resemble the structure of the query sequence, and on the production of an alignment that maps residues in the query sequence to residues in the template sequence.

- It has been shown that protein structures are more conserved than protein sequences amongst homologues, but sequences falling below a 20% sequence identity can have very different structure.

- Evolutionarily related proteins have similar sequences and naturally occurring homologous proteins have similar protein structure. It has been shown that three-dimensional protein structure is evolutionarily more conserved than would be expected on the basis of sequence conservation alone.

- The sequence alignment and template structure are then used to produce a structural model of the target. Because protein structures are more conserved than DNA sequences, detectable levels of sequence similarity usually imply significant structural similarity.

Genome sequence annotation-:

-DNA annotation or **genome annotation** is the process of identifying the locations of genes and all of the coding regions in a genome and determining what those genes do.

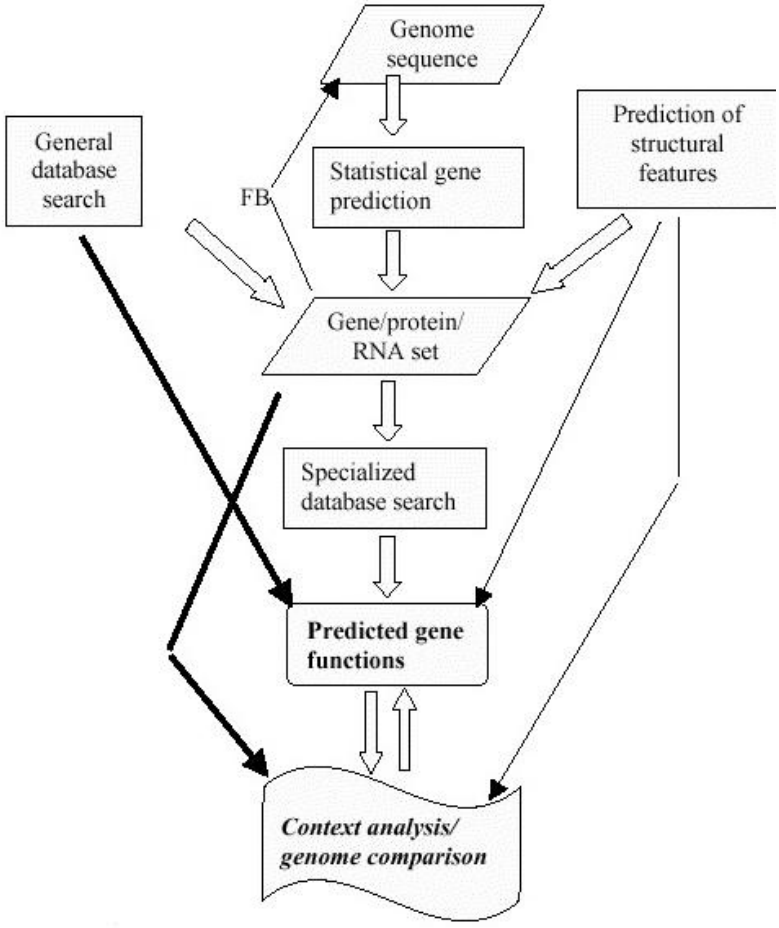
-An annotation (irrespective of the context) is a note added by way of explanation or commentary. Once a genome is sequenced, it needs to be annotated to make sense of it.

-For DNA annotation, a previously unknown sequence representation of genetic material is enriched with information relating genomic position to intron-exon boundaries, regulatory sequences, repeats, gene names and protein products.

- This annotation is stored in genomic databases such as Mouse Genome Informatics, FlyBase, and WormBase.

Steps of genome sequence annotation-

.



ORF identification-:

- DNA is a genetic material that contains all the genetic information in a living organism. The information is stored at genetic code using ATGC. During the transcription process DNA is transcribed to mRNA.
- Each of theses base pair will bond with a sugar and phosphate molecule to form a nucleotide. This nucleotide that codes for a particular amino acids. During translation is a CODON.

- The region of a nucleotide that starts from Initiation codon and ends with a stop codon is called as ORF (open reading frame).
- Proteins are formed from ORF by analyzing the ORF we can predict the possible amino acids that might be produced during translation.
- The ORF finder is a program that is available at NCBI website it identifies all ORF or protein coding regions from 6 different frames.

HOW TO FIND ORF?

- Consider a hypothetical sequence.

  Ex- CAT, GGA, GTA, TCG, CAG, GGT, CAA. (First reading frame).

- The second reading is formed after leaving the first one nucleotide and then grouping the sequences into words of 3 nucleotide.

  C ATGGAGTATCGCAGGGTC AA (Second reading frame).

- The third reading frame is formed after leaving the first 2 nucleotides and then grouping the sequences into words of 3 nucleotides.

  CA TGGAGTATCGCAGGGTCA A (Third reading frame).

- Others reading frame will be obtained by just finding the reverse complementary sequence of the above reading frames.

  TTGACCCTGCGATACTCCATG (fourth)

  T TGACCCTGCGATACTCCA TG (fifth)

  TT GACCCTGCGATACTCCAT G (sixth)

- In all the reading frames formed, we have to replace thymine with uracil (T with U) , now mark the start and stop codon in the reading frame.

  CAUGGAGUAUCGCAG GGT CAA.

  C AUGGAGUAU CGCAGGGUC AA

  CA UGGAGUAUCGCAGGGUCAA

  UUGACCCUGCGAUACUCCAUG

  U UGACCCUGCGAUACUCCAUG

  UUGACCCUGCGAUACUCCAU G

Start codons – AUG

Stop codons UAA, UGA, UAG.

Biological network and its identification

A **biological network** is any network that applies to biological systems. A network is any system with sub-units that are linked into a whole, such as species units linked into a whole food web. Biological networks provide a mathematical representation of connections found in ecological, evolutionary, and physiological studies, such as neural networks. The analysis of biological networks with respect to human diseases has led to the field of network medicine.

Protein–protein interaction networks

Many protein–protein interactions (PPIs) in a cell form protein interaction networks (PINs) where proteins are nodes and their interactions are edges. PINs are the most intensely analyzed networks in biology. There are dozens of PPI detection methods to identify such interactions. The yeast two-hybrid system is a commonly used experimental technique for the study of binary interactions.

Recent studies have indicated conservation of molecular networks through deep evolutionary time. Moreover, it has been discovered that proteins with high degrees of connectedness are more likely to be essential for survival than proteins with lesser degrees. This suggests that the overall composition of the network (not simply interactions between protein pairs) is important for the overall functioning of an organism.

Gene regulatory networks (DNA–protein interaction networks)

The activity of genes is regulated by transcription factors, proteins that typically bind to DNA. Most transcription factors bind to multiple binding sites in a genome. As a result, all cells have complex gene regulatory networks. For instance, the human genome encodes on the order of 1,400 DNA-binding transcription factors that regulate the expression of more than 20,000 human genes.

Gene co-expression networks (transcript–transcript association networks)

Gene co-expression networks can be interpreted as association networks between variables that measure transcript abundances. These networks have been used to provide a systems biologic analysis of DNA microarray data, RNA sequence data, miRNA data etc. weighted gene co-expression network analysis is widely used to identify co-expression modules and intramodular hub genes. Co-expression modules may correspond to cell types or pathways. Highly connected intramodular hubs can be interpreted as representatives of their respective module.

Metabolic networks

The chemical compounds of a living cell are connected by biochemical reactions which convert one compound into another. The reactions are catalyzed by enzymes. Thus, all compounds in a

cell are parts of an intricate biochemical network of reactions which is called metabolic network. It is possible to use network analyses to infer how selection acts on metabolic pathways.

Signaling networks

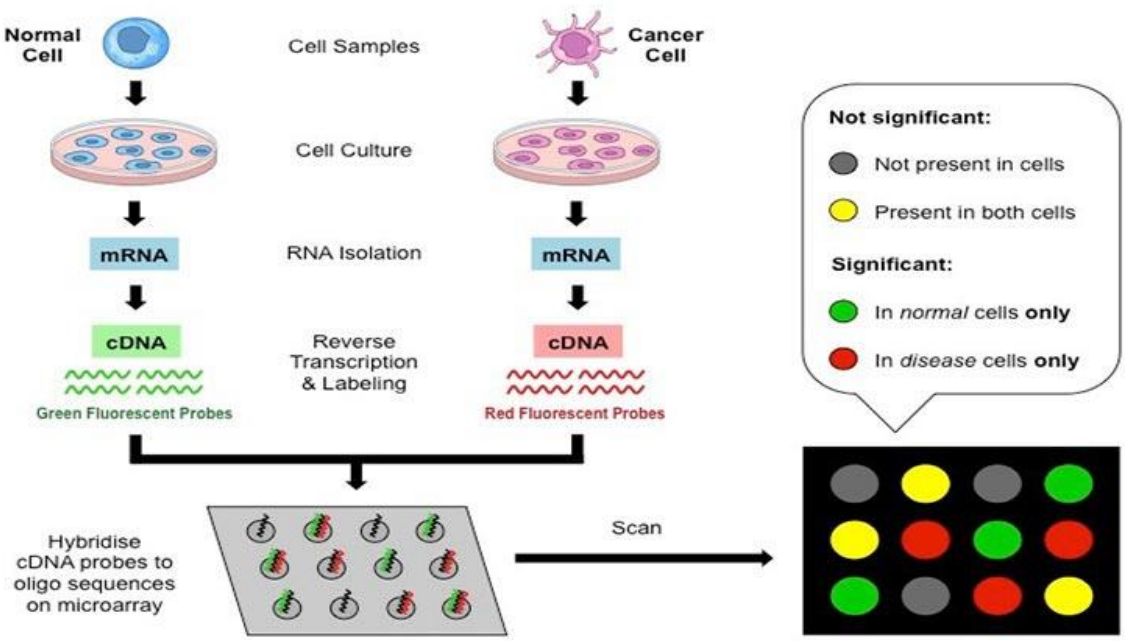
Signals are transduced within cells or in between cells and thus form complex signaling networks. For instance, in the MAPK/ERK pathway is transduced from the cell surface to the cell nucleus by a series of protein–protein interactions, phosphorylation reactions, and other events. Signaling networks typically integrate protein–protein interaction networks, gene regulatory networks, and metabolic networks.

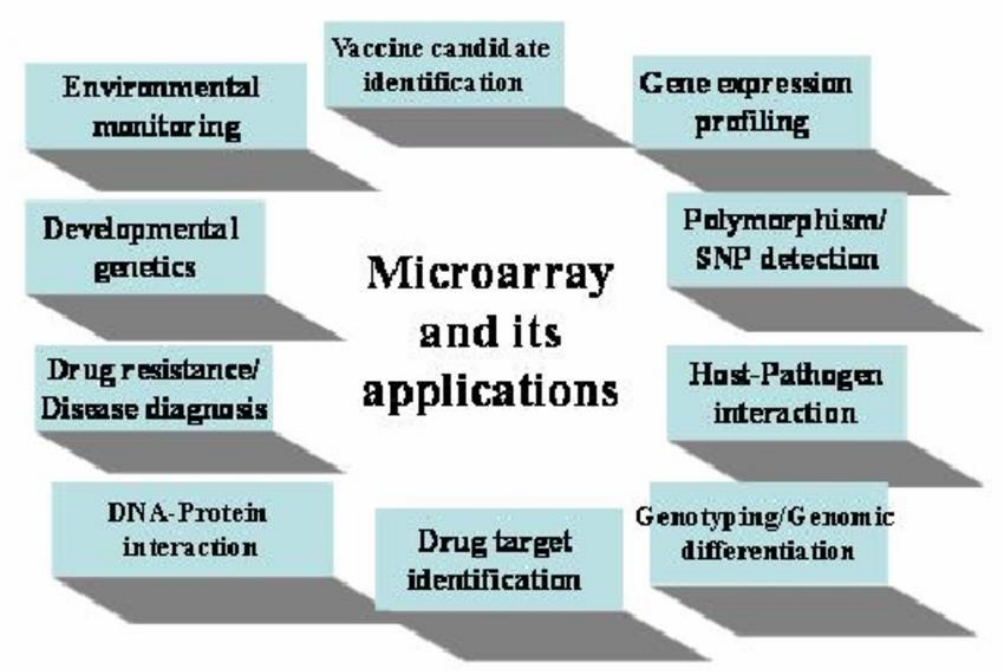
Neuronal networks

The complex interactions in the brain make it a perfect candidate to apply network theory. Neurons in the brain are deeply connected with one another and this results in complex networks being present in the structural and functional aspects of the brain. For instance, small-world network properties have been demonstrated in connections between cortical areas of the primate brain or during swallowing in humans. This suggests that cortical areas of the brain are not directly interacting with each other, but most areas can be reached from all others through only a few interactions.

Microarray technique

Microarray analysis techniques are used in interpreting the data generated from experiments on DNA (**Gene chip analysis**), RNA, and protein microarrays, which allow researchers to investigate the expression state of a large number of genes - in many cases, an organism's entire genome - in a single experiment.Such experiments can generate very large amounts of data, allowing researchers to assess the overall state of a cell or organism. Data in such large quantities is difficult - if not impossible - to analyze without the help of computer programs.





Protein function prediction

- Protein function prediction are method or technique that bioinformatics researcher used to assign biochemical roles to proteins. These proteins are usually once that are poorly studied or predicted based on genomic sequence data. These predictions are often driven by data intensive computational procedure.

- Information may come from nucleic acids sequence homology, gene expression profile, protein domain structure, mining of publication , phylogenetic profile , protein-protein interactions.

Methods-

1- Homology based method-

- Proteins of similar sequences are usually homologous and thus have similar functions.
- Hence protein in a newly sequence genome are properly annotated using the sequences of similar in relation to genome.
- However closely related protein do not always share same function.
  For ex- the yeast GAL1 and GAL3 paralogs (73% identical & 92% similar) have evolved vary in different function with GAL1 being an (Galactokinase) andGAL3 being an transcription inducer.

2- Sequence motif based method-

- The development of protein database such as pFam (protein families database) allow us to find known domains within a query sequence providing evidence for likely functions.
- Within protein domains shorter signature are known as Motif.
- Motifs are associated with particular function and motif databases such as PROSITE (database of protein domain families and functional sites) can be search using a query sequence.

3- Structure based method-

- 3D protein structure is generally well conserved than protein sequences, structure similarity is a good indicator of similar function in 2 or more proteins . Many programs have been developed to screen and find unknown protein structural and its function against protein database and to find similarity in their structure as well as function.

4- **Genomic context based method-**

Many of the never method for protein function prediction are not based on the comparison of sequence or structure as above.

But on some type of correlation between novel gene or protein and though that already have annotation. It is also knowns as Phylogenomic profiling.

5- Network based method-

These are based on different data sources which can be combine into a composite network and then we used by prediction algorithm to annotate genes and proteins. Tool based on this method is STRING.

Next generation sequencing (NGS)

-
- NGS massively parallel or deep sequencing are related terms that describes a DNA sequencing technology which have revolutionised Genomic research.
- Using NGS and entire human genome can be sequenced within a single day. In contrast the previous Sanger sequencing technology used to sequence human genome over a decade.
- There are no. of different NGS platforms. All platforms performs sequencing of millions of small fragments of DNA inputs.

Potential uses of NGS in clinical practice:-

1- NGS captures a broader spectrum of mutations than sanger sequencing-:

- The spectrum of DNA variations in a human genome comprises of small base changes, insertions and deletions of DNA, large genomic deletions of exons or whole genes and rearrangement such as inversions and translocations.
- Traditional Sanger sequencing is restricted to the discovery of substitutions and small insertions and deletions.
- For the remaining mutations dedicated assays are frequently performed such as Fluorescence in situ hybridization (FISH) for conventional karyotyping or comparative genomic hybridization microarrays to detect submicroscopic chromosomal copy number changes such as microdeletions.
- However, these data can also be derived from NGS sequencing data directly obviating the need for dedicated assays while harvesting the full spectrum of genomic variation in single experiment.
- Limitations:- in regions which sequence poorly or map erroneously due to extreme guanine or cytosine content or repeat architecture.
  Ex- The repeat expansions under lying fragile X syndrome or Huntington's disease.

2- Genomes can be interrogated without bias-:

A). **Microbiology-:**

- The main ability of NGS in microbiology is to replace conventional characterization of pathogens by morphology, staining properties and metabolic criteria with a genomic definition of pathogens.

- When NGS was used to reveal and trace an outbreak of Methicillin- resistant staphylococcus aureus(MRSA) on a neonatal intensive care unit.

B).Oncology:-

The fundamental premise of cancer genomics is that cancer is caused by somatically acquired mutations, and consequently it is a disease of the genome .

Although capillary based cancer sequencing has been ongoing for over a decade .

Currently pilot project are underway using NGS of cancer genomes in clinical practice ,mainly aiming to identify mutation in tumour that can be targeted by mutation specific drugs.

Limitation:-

The main disadvantages of NGS in the clinical setting is putting in plce the required infrastructure such as computer capacity and storage , and also the personel expertise required to comprehensively analyse and interpret the subsequent data.

The actual sequencing cost of NGS is negligibile.

NGS has huge potential but is presently used primarily for research.

NGS will allow paediatricians to take genetic information to the beside.

PROTEIN STRUCTURE PREDICTION (SECONDARY AND TERTIARY STRUCTURE).

(A) SECONDARY STRUCTURE PREDICTION

INTRODUCTION

In Secondary Structure Prediction, We will get three dimensional structure of protein, from that three dimensional structure we will get the function of the specific protein.

Secondary Structure Prediction are devised by many tools.

PURPOSE OF SECONDARY STRUCTURE PREDICTION /OBJECTIVES

- Identifies signal peptide.
- Describe the alpha helices and beta sheets.
- Predicts the location of membrane spanning helices.
- Identifies membrane spanning beta sheets.
- Predicts the secondary structure from linear sequence.
- Analyze results.
- Performs 3D structure via homology modelling.

SECONDARY STRUCUTURE ELEMENTS

ALPHA HELIX

- Most abundant secondary structure.
- It has 3.6 amino acids per turn.
- Average length have 10 amino acids, it varies from5 to 40.
- Inner facing side chains are hydrophobic.
- Third of every fourth amino acid is hydrophobic.

BETA SHEET

- Hydrogen bond between two separate regions of chain.
- Each ~5 to 10 amino acid at each region form beta sheets.
- There are of Parallel – Same Direction
- Anti- parallel- Different Direction

METHODS OF SECONDARY STRUCTURE PREDICTION

- Chou- Fasman methods
- GOR methods
- Nearest neighbour methods
- Hidden Markov models
- Neural networks
- Multiple alignment based self optimization method

(1) CHOU FASMAN METHOD

- In Chou Fasman method, the propensity value is important
- R- group attached to the protein chain are responsible for the propensity value.
- The main terms used in Chou Fasman method was Alpha helix or Beta sheet makers , Alpha helix or Beta sheet breakers

Propensity Value

Many online and offline server tools are available to predict the secondary structure by Chou Fasman method.

ALPHA HELIX MAKERS:

Alanine

Glutamine

Leucine

Methionine

ALPHA HELIX BREAKERS:

Proline Glysine

BETA SHEET MAKERS:

Isoleucine

Valine

Tyrosine

BETA SHEET BREAKERS:

Proline

Asparagine Glutamine

PROPENSITY VALUE

- Tendency of the aminoacids to behave more in alpha helix and beta sheet.
- Propensity value for Alpha Helix = Frequency of amino acids in Alpha helix Frequency of residues to be in Alpha helix
- If Alpha helix is made up of 20 amino acids and amino acids present for 5 times , Frequency of amino acids = 5/20= 0.25
- If the total 100 residues in the protein, but only 20 makes the Alpha helix, Frequency of residues to be in the helix =20/100=0.2
- This is applicable for beta sheets also.

Rule for Alpha helix formation:

- Firstly, we have to scan 6 amino acid residues at random.
- Then look for maker and breakers
- If there is more than 1/3 of breaker, there is no helix formation.
- If there is less than 1/2 of makers, there is no helix formation.
- Naturally it adding the stretch of sequence till the propensity value of helix is greater than OR equal to 100.

- Four such amino acid have propensity value of helix is greater than 100, it terminates the elongation.
- There are 9 amino acid sequence with more maker and less breaker make Alpha helix.
- Negative charged amino acid at N terminal and Positive charged amino acid at C terminal form the ALPHA HELIX.
- If propensity value for Alpha helix is greater than OR equal to 1.03, form the ALPHA HELIX.
- Alpha helix makers should be greater than Alpha helix breakers.
- The propensity value of Alpha helix should be greater than the propensity value of Beta sheet.

Rules for Beta Sheet formation:

- Firstly, we have to scan 5 amino acid residues at random.
- If propensity value for Beta sheet is greater than OR equal to 1.05, form the BETA SHEET.
- Other rules are as same as Alpha helix formation.

(2) GOR METHOD

- GOR method assumes that amino acids up to 8 residues on each side influence the secondary structure of the central residue.
- This program is now fourth version.
- The accuracy of GOR when checked against a set of 267 proteins of known structure is 64%.
- This implies that 64% of the amino acids were correctly predicted as being helix, sheet or coil.
- The algorithm uses a sliding window of 17 amino acids.

(3) NEAREST NEIGHBOUR METHOD

- It is based on the hypothesis that short homologous sequences of amino acids have the same secondary structure tendencies.
- A list of short sequence s is made by sliding a window of length n along a set of approximately 100- 400 training sequences of known structure but minimal sequence similarity.

(4)HIDDEN MARKOV METHOD

- HIDDEN MARKOV means the state is directly invisible to be observer.
- It provide a basic description of what pfam provides.
- Each HMM is trained with the sequences of the protein in that structural class.
- The models used are with a query sequence to predict both the class and secondary structure.

(5)NEURAL NETWORKS

- Most effective structure prediction tool for pattern recognition and classification.
- The protein sequence is translated into patterns by shifting a window of n adjacent residues (n= 13-21) through the protein.

Methods to predict,

Hierchical Neural Network

nnPredict
PSA
PSIPRED
Gen THREADER
MEMSAT
PSI BLAST Version 2.0

(6) MULTIPLE ALIGNMENTS BASED SELF- OPTIMIZATION METHOD

- SOPMA correctly predicts 69.5% of amino acids for a three state description of the secondary structure in a whole database containing 126 chains of nonhomologous proteins.
- Joint prediction with SOPMA and PHD correctly predicts 82.2% of residues for 74% of co- predicted amino acids.

(2) TERTIARY STRUCTURE PREDICTION

Introduction

- Protein three-dimensional structures are obtained using two popular experimental techniques, x-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy.
- There are many important proteins for which the sequence information is available, but their threedimensional structures remain unknown.
- Therefore, it is often necessary to obtain approximate protein structures through computer modeling.
- Having a computer-generated threedimensional model of a protein of interest has many ramifications, assuming it is reasonably correct.
- It may be of use for the rational design of biochemical experiments, such as site-directed mutagenesis, protein stability, or functional analysis.
- There are three computational approaches to protein three-dimensional structural modeling and prediction.
- They are **homology modeling, threading, and ab initio prediction.**
- The first two are knowledge-based methods; they predict protein structures based on knowledge of existing protein structural information in databases.
- The ab initio approach is simulation based and predicts structures based on physicochemical principles governing protein folding without the use of structural templates.

Homology modelling

- As the name suggests, homology modeling predicts protein structures based on sequence homology with known structures.
- It is also known as comparative modeling.
- The principle behind it is that if two proteins share a high enough sequence similarity, they are likely to have very similar three-dimensional structures.
- If one of the protein sequences has a known structure, then the structure can be copied to the unknown protein with a high degree of confidence.
- The overall homology modeling procedure consists of six major steps and one additional step.

1**. Template Selection** :-

- The template selection involves searching the Protein Data Bank (PDB) for homologous proteins with determined structures.
- The search can be performed using a heuristic pairwise alignment search program such as BLAST or FASTA.
- However, programming based search programmes such as SSEARCH or ScanPS can result in more sensitive search results.
- Homology models are classified into 3 areas in terms of their accuracy and reliability.
- Midnight Zone: Less than 20% sequence identity. The structure cannot reliably be used as a template.
- Twilight Zone: 20% - 40% sequence identity.
- Sequence identity may imply structural identity.
- Safe Zone: 40% or more sequence identity. It is very likely that sequence identity implies structural identity
- Often, multiple homologous sequences may be found in the database. Then the sequence with the highest homology must be used as the template.

2. Sequence Alignment:

- Once the structure with the highest sequence similarity is identified as a template, the full-length sequences of the template and target proteins need to be realigned using refined alignment algorithms to obtain optimal alignment.
- Incorrect alignment at this stage leads to incorrect designation of homologous residues and therefore to incorrect structural models.
- Therefore, the best possible multiple alignment algorithms, such as Praline and T-Coffee should be used for this purpose.

3. Backbone Model Building:

- Once optimal alignment is achieved, the coordinates of the corresponding residues of the template proteins can be simply copied onto the target protein.
- If the two aligned residues are identical, coordinates of the side chain atoms are copied along with the main chain atoms.
- If the two residues differ, only the backbone atoms can be copied.

4. Loop Modeling :

- In the sequence alignment for modeling, there are often regions caused by insertions and deletions producing gaps in sequence alignment.
- The gaps cannot be directly modeled, creating "holes" in the model.
- Closing the gaps requires loop modeling which is a very difficult problem in homology modeling and is also a major source of error.
- Currently, there are two main techniques used to approach the problem: the database searching method and the ab initio method.
- The database method involves finding "spare parts" from known protein structures in a database that fit onto the two stem regions of the target protein.
- The stems are defined as the main chain atoms that precede and follow the loop to be modeled.
- The best loop can be selected based on sequence similarity as well as minimal steric clashes with the neighboring parts of the structure.
- The conformation of the best matching fragments is then copied onto the anchoring points of the stems.
- The ab initio method generates many random loops and searches for the one that does not clash with nearby side chains and also has reasonably low energy and φ and ψ angles in the allowable regions in the Ramachandran plot.
- Schematic of loop modeling by fitting a loop structure onto the endpoints of existing stem structures represented by cylinders.
- FREAD is a web server that models loops using the database approach.
- PETRA is a web server that uses the ab initio method to model loops.
- CODA is a web server that uses a consensus method based on the prediction results from FREAD and PETRA.

5. Side Chain Refinement:

- Once main chain atoms are built, the positions of side chains that are not modeled must be determined.
- A side chain can be built by searching every possible conformation at every torsion angle of the side chain to select the one that has the lowest interaction energy with neighboring atoms.
- Most current side chain prediction programs use the concept of rotamers, which are favored side chain torsion angles extracted from known protein crystal structures.
- A collection of preferred side chain conformations is a rotamer library in which the rotamers are ranked by their frequency of occurrence.
- In prediction of side chain conformation, only the possible rotamers with the lowest interaction energy with nearby atoms are selected.

- A specialized side chain modeling program that has reasonably good performance is SCWRL, which is a UNIX program.

6. Model Refinement :

- In these loop modeling and side chain modeling steps, potential energy calculations are applied to improve the model.
- Modeling often produces unfavorable bond lengths, bond angles, torsion angles and contacts.
- Therefore, it is important to minimize energy to regularize local bond and angle geometry and to relax close contacts and geometric chain.
- The goal of energy minimization is to relieve steric collisions and strains without significantly altering the overall structure.
- However, energy minimization has to be used with caution because excessive energy minimization often moves residues away from their correct positions.
- GROMOS is a UNIX program for molecular dynamic simulation. It is capable of performing energy minimization and thermodynamic simulation of proteins, nucleic acids, and other biological macromolecules.
- The simulation can be done in vacuum or in solvents.
- A lightweight version of GROMOS has been incorporated in SwissPDB Viewer.

7. Model Evaluation:

- The final homology model has to be evaluated to make sure that the structural features of the model are consistent with the physicochemical rules.
- This involves checking anomalies in φ–ψ angles, bond lengths, close contacts, and so on.
- If structural irregularities are found, the region is considered to have errors and has to be further refined.
- Procheck is a UNIX program that is able to check general physicochemical parameters such as φ–ψ angles, chirality, bond lengths, bond angles, and so on.
- WHAT IF is a comprehensive protein analysis server that has many functions, including checking of planarity, collisions with symmetry axes, proline puckering, anomalous bond angles, and bond lengths.
- Few other programs for this step are ANOLEA, Verify3D, ERRAT, WHATCHECK, SOV etc.

Threading/Fold recognition

- By definition, threading or structural fold recognition predicts the structural fold of an unknown protein sequence by fitting the sequence into a structural database and selecting the best-fitting fold.

- The comparison emphasizes matching of secondary structures, which are most evolutionarily conserved.
- The algorithms can be classified into two categories, pairwise energy based and profile based.

Pairwise Energy Method

- In the pairwise energy based method, a protein sequence is searched for in a structural fold database to find the best matching structural fold using energy-based criteria.
- The detailed procedure involves aligning the query sequence with each structural fold in a fold library.
- The alignment is performed essentially at the sequence profile level using dynamic programming or heuristic approaches.
- Local alignment is often adjusted to get lower energy and thus better fitting.
- The next step is to build a crude model for the target sequence by replacing aligned residues in the template structure with the corresponding residues in the query.
- The third step is to calculate the energy terms of the raw model, which include pairwise residue interaction energy, solvation energy, and hydrophobic energy.
- Finally, the models are ranked based on the energy terms to find the lowest energy fold that corresponds to the structurally most compatible fold.

Profile Method

- In the profile-based method, a profile is constructed for a group of related protein structures.
- The structural profile is generated by superimposition of the structures to expose corresponding residues.
- Statistical information from these aligned residues is then used to construct a profile.
- The profile contains scores that describe the propensity of each of the twenty amino acid residues to be at each profile position.
- To predict the structural fold of an unknown query sequence, the query sequence is first predicted for its secondary structure, solvent accessibility, and polarity.
- The predicted information is then used for comparison with propensity profiles of known structural folds to find the fold that best represents the predicted profile.
- Threading and fold recognition assess the compatibility of an amino acid sequence with a known structure in a fold library.

- If the protein fold to be predicted does not exist in the fold library, the method will fail.
- 3D-PSSM, GenThreader, Fugue are few web based programmes used for threading.

Ab initio method

- When no suitable structure templates can be found, Ab Initio methods can be used to predict the protein structure from the sequence information only.
- As the name suggests, the ab initio prediction method attempts to produce allatom protein models based on sequence information alone without the aid of known protein structures.
- Protein folding is modeled based on global free-energy minimization.
- Since the protein folding problem has not yet been solved, the ab initio prediction methods are still experimental and can be quite unreliable.
- One of the top ab initio prediction methods is called Rosetta, which was found to be able to successfully predict 61% of structures (80 of 131) within 6.0 Å RMSD (Bonneau et al., 2002).

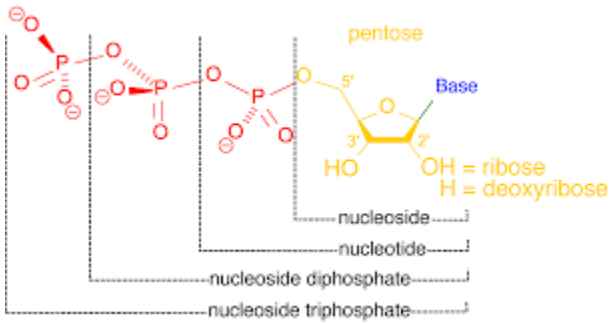
UNIT-2

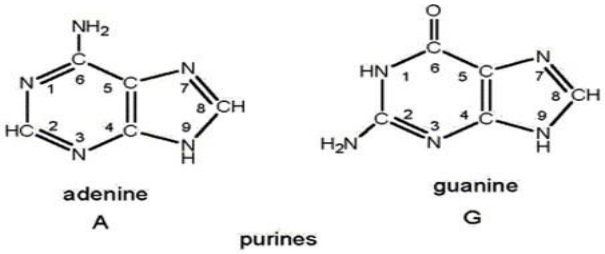
BASIC CONCEPTS OF RNA SECONDARY STRUCTURE PREDICTION

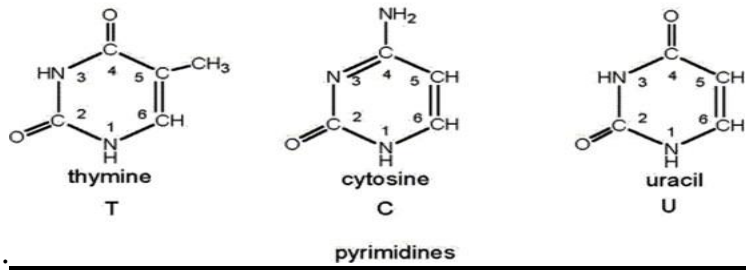
What is RNA and where it is used?

1- Proteins and nucleic acids like RNA &DNA play an important role in reproducing and maintaining life.
2- Proteins are important because they control processes like energy metabolism,intercellular communication and biosynthesis. They synthesized using the genetic information which is stored in a DNA.
3- RNA molecules are used for the synthesis of proteins, the act as messenger.
4- Both DNA & RNA are composed of subunit, so called nucleotide or bases.
5- There are only four different type nucleotide in a molecule but DNA & RNA do have other sets of nucleotide.
6- Nucleotides consist of a nitrogen containing base, a 5-carbon sugar ring & a phosphate group.
7- The nucleotides are linked together by phosphodiester linkage through the hydroxyl group on the sugar on one nucleotide and a phosphate on the next one.
8- As a result one can observe a stand with the so called 5 prime end (5'end) where a free phosphate group can be formed and the 3' end with a free hydroxyl group.

9-The nitrogenous base has the structure of a planar ring & is either purine or pyrimidine.

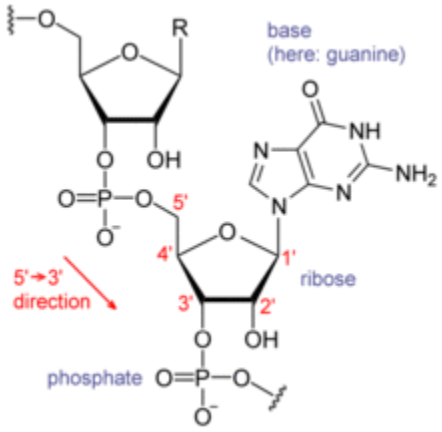






THE STRUCTURE OF RNA-

1- In RNA, nucleotides the sugar which is used is ribose and therefore they are also called ribonucleotide.
2- The purine bases are adenine and guanine but the pyrimidines are cytosine and uracil.
3- RNA molecules are much smaller than DNA molecules & they are also called as linear polymers.
4- Moreover they do not seen to have regular 3-D structure and are mostly single standard.
5- This makes them more flexible than DNA & they can also act as enzymes.
6- The molecules also contain a very stable 3-D structure with unpaired region which are very flexible.
7- The wobble base pair makes an important factor for this flexibility.
8- Beside the WATSON-CRICK BASE PAIR, A:U & G:C , is the wobble base pair G:U one of the most common base pair in RNA molecule.
9- But actually, any of the bases can build hydrogen bond with any other base.
10- Another difference between DNA & RNA is that double stranded RNA builds alpha-helices while ds DNA builds beta-helices.
11- The major group of the alpha-helix is rather narrow and bigger .This is due to ribose needing more space than deoxyribose.



PRIMARY, SECONDARY& TERTIARY STRUCTURE -

1-The primary structure of a molecule describe only the 1-D sequence of its components.
2-The primary structure of RNA is almost identical to the primary structure of DNA besides the component being A,C,G&U. instead of T.
3-The secondary structure of molecules is more complex than the primary structure and can been drawn in 2-D space.
4- RNA secondary structure is mainly composed of double stranded RNA regions form by folding the single stranded RNA molecule back on itself.
5-The tertiary structure is an overall 3-D structure of molecules.
6-It is build on the interaction of the lower border secondary structure.
7-Helices are examples of RNA & DNA tertiary structure.

8-Pseudoknot is a tertiary structure of RNA.

****IMPORTANCE OF RNA SECONDARY STRUCTURE PREDICTION-**

1-Important aspect of the prediction of RNA secondary structure is that the there are many sequences whose structures have not yet been experimentally determined &for which there are no homologs in the databases from which the structure could be deriving. Hence, it is a good idea to predict the structure.

2-Moreover,it has been shown that RNA secondary structure prediction has application to design of nucleic acid probes.

3- It is also used by molecular biologist to help& predict conserved structural element in non-coding region of gene transcripts.

4-Finally, there is also application in predicting structure that are conserved during evolution.

Different secondary structure elements-:

1-STEM LOOP (HAIRPIN LOOPS).

2-BULGE LOOPS.

3-INTERIOR LOOPS.

4-JUNCTION OR MULTI LOOPS.

5-PSEUDOKNOTS.

DESCRIPTION OF THESE ABOVE ONES-

1- **Stem loops or hair pin loop-** stem loops is a lollipop shape structure form when a single stranded nucleic acid molecules loops back on itself to form a complementary double helix topped by a loop. They are at least four bases long.

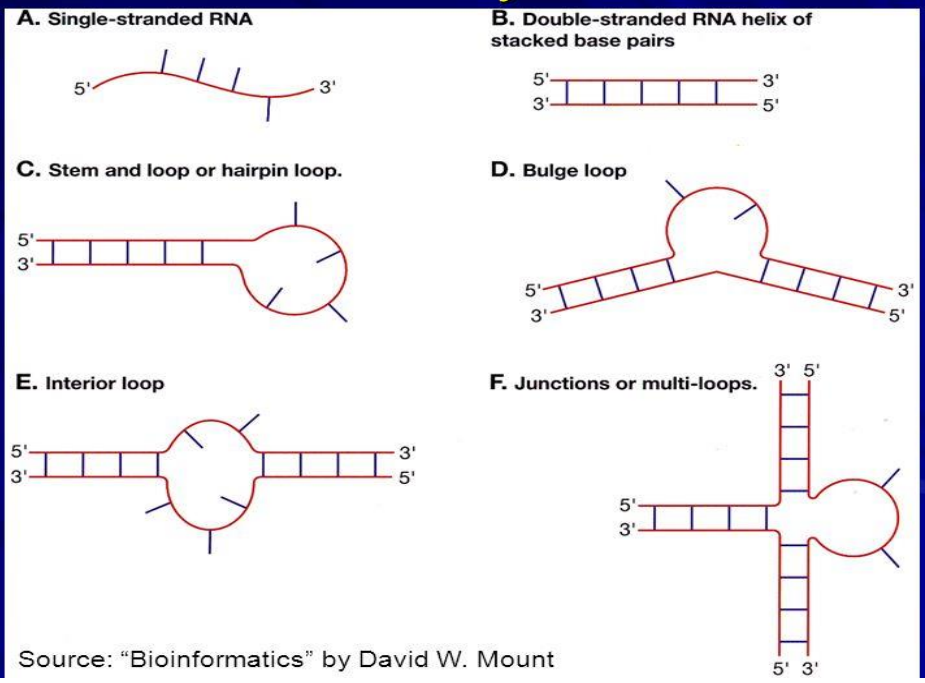
2-Bulge loops-Bulge loops are commonly found in helical segments of cellular RNAs. Bulge loops occurs when bases on one side of the structure cannot form this pair & they cause bends in the helix.

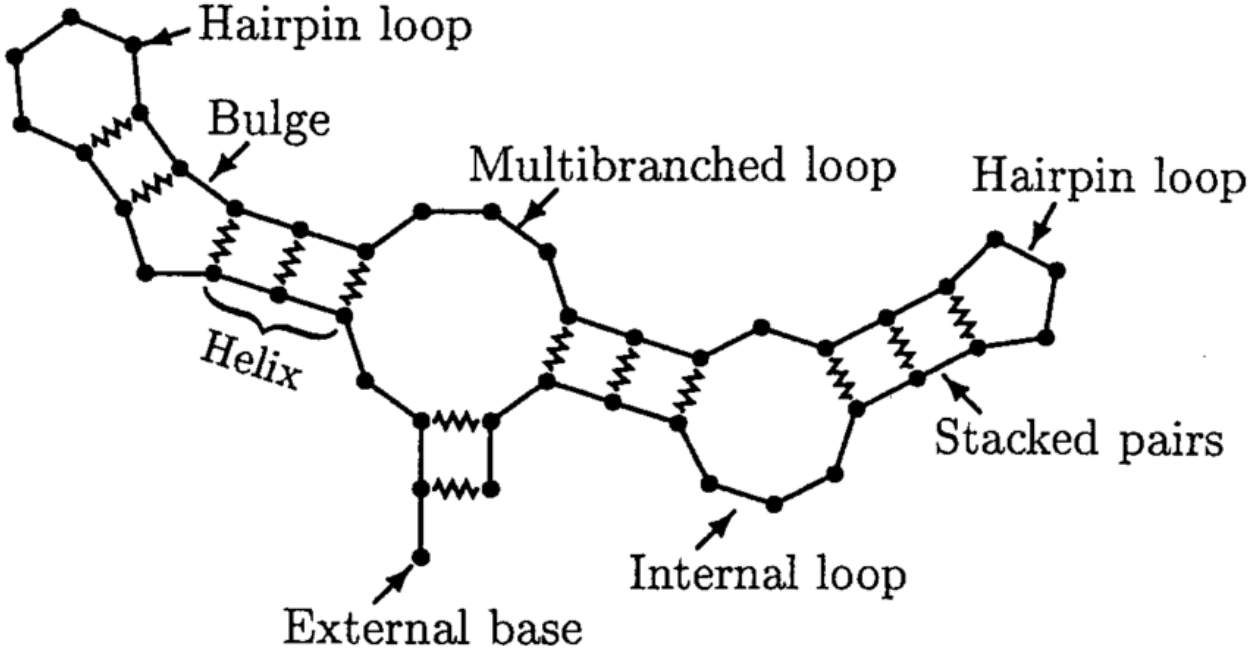
3-Interior loops-Interior loops occur when bases on both sides of the structure cannot form base pair.

4-Junction or multi loop-Junction in loop 2 or more double stranded region conversing to form a closed structure.

5-Pseudoknots- Pseudoknots is a tertiary structural elements of RNA. It is formed by base pairing between an already existing secondary structure loops & the free ending. Nucleotides within a hair pin loop, form base pairs with nucleotides outside the stem. Hence, this pair occurs that overlap each other in their sequence position.

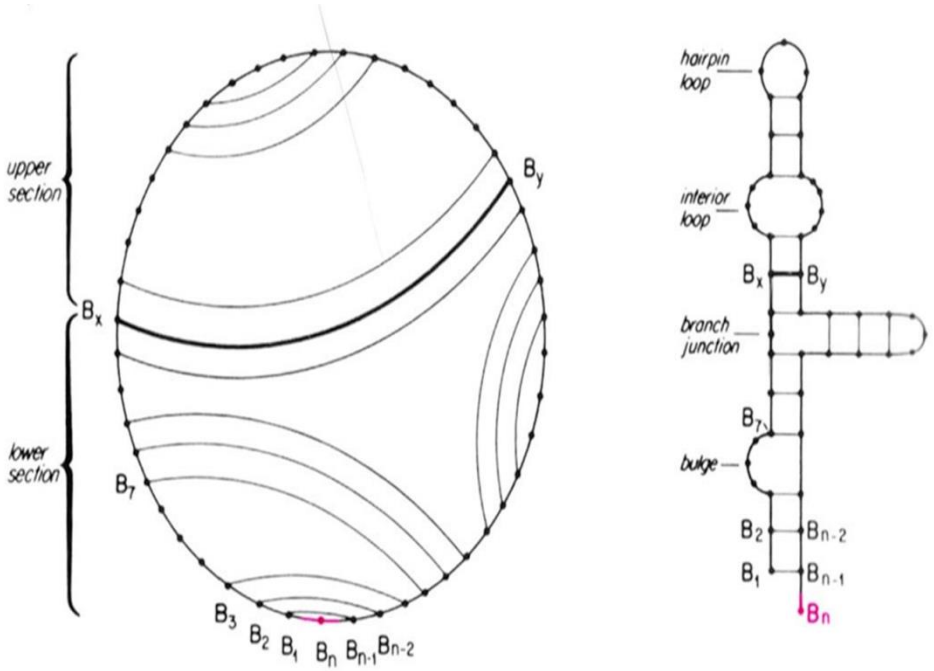
Types of single- & double-stranded regions in RNA secondary structures.





ASSUMPTIONS IN RNA STRUCTURE PREDICTION

- The most likely structure is similar to the energetically most stable structure.
- The energy associated with any position in the structure is only influenced by local sequence and strtucture.
- The structure formed does not produce pseudoknots.
- One method of representing the base pairs of a secondary structure is to draw the structure in a circle.
- An arc is drawn to represent each base pairing found in the structure.
- If any of the arc cross, then a pseudoknot is present.



RNA structure prediction methods

- Base Pair Maximization
- Energy Minimization

Base Pairs Maximization

- This approach is to find the configuration with the greatest numbers of paired bases.
- Given a RNA sequence, determine the set of maximal base pairs(no base pair across each other)

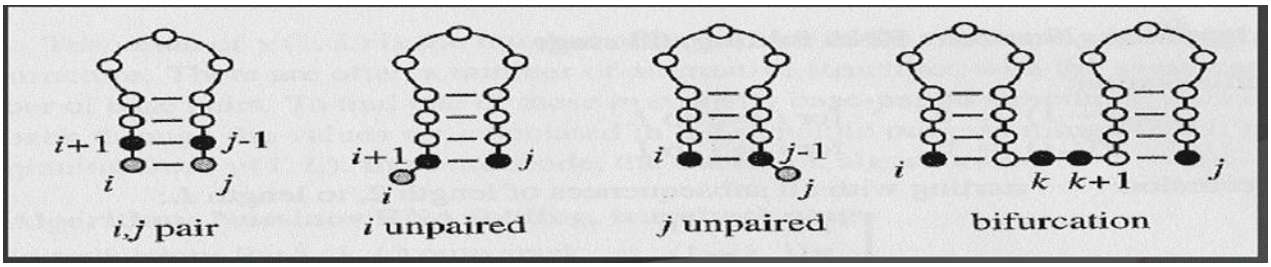
• Align bases according to their ability to pair with each other gives an approach to determining the optimal structure

Methods adopted

- Dynamic programming approach
- Nussinov Algorithm

Nussinov Algorithm

Four ways to get the optimal structure between position I and j from the optimal substructure
1.Add i,j pair onto best structure found for subsequence i+1,j-1
2.Add unpaired position i onto best structure for subsequence i +1,j
3.Add unpaired position i onto best structure for subsequence I, j-1
4.Combine two optimal structures i,k and k+1,j



NussinovAlgorithm

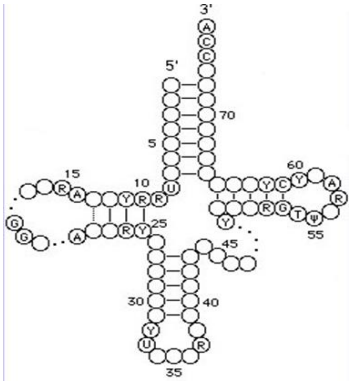
- compares a sequence against itself in a n*n matrix
- Find the maximum of the scores for the four possible structures at a particular position.

Base Pair Maximization - Drawbacks

- Base pair maximization will not necessarily lead to the most stable structure
- May create structure with many interior loops or hairpins which are energetically unfavorable
- Comparable to aligning sequences with scattered matches – not biologically reasonable

Energy Minimization

- Thermodynamic Stability
- Estimated using experimental techniques
- Theory : Most Stable is the Most likely
- No Pseudknots due to algorithm limitations
- Uses Dynamic Programming alignment technique



• Attempts to maximize the score taking into account thermodynamics
- Gaps represent some form of a loop
- The most widely used software that incorporates this minimum free energy algorithm is MFOLD/RNAfold/ViennaRNA

Energy Minimization Drawbacks

- Compute only one optimal structure
- Usual drawbacks of purely mathematical approaches
- Similar difficulties in other algorithms
- Protein structure
- Exon finding

Suboptimal structure prediction by M-fold:-

- M-fold predicts optimal and suboptimal secondary structures for RNA or DNA molecule using the most recent energy minimization method.
- M-fold is an adaptation of the M-fold package (version 2.3) by zuker and taeqer that has been modifying to work with the Wisconsin package.these method uses and energy rues develop by turner and colleagues to determine optimal and sub-optimal secondary structure for an RNA molecule.
- Using energy minimization criteria any predicted optimal secondary structure for an RNA or DNA molecules depends on model of folding and specific folding energies used to calculate the structure.

Covariance Method:-

- It describe a general approach to several RNA sequence analysis problems using probabilistics models that flexibly describe the secondary structure ad primary sequence consensus of an RNA sequence family.
  These models is known as covariance models.
- A covariance model of t RNA sequence is an extremely sensitive and discriminative tool for searching for additional t RNA and tRNA related sequences in sequence database.
- A model can be automatically from an existing sequence alignment.
- Also describe an algorithm for learning a model and hence a consensus secondary structure from initially unaligned example sequence and no prior structural information.

  Models trained on unaligned t RNA example correctly predict t RNA secondary structure and produce high quality multiple alignment.
  The approach may be applied to any family of small RNA sequences.
- Probabilistic model, also known as covariance model ,which cleanly describes both secondary structure and primary sequence consensus of an RNA .
- Using covariance models ,we introduce new and general approaches to several RNA analysis problems: consensus secondary structure prediction,,ultiple sequence alignment and databse similarity searching.
  Also describe a dynamic programming algorithm for efficiently finding the globally optimal alignments of RNA sequences to a model , and show how to use the algorithm for database searching.

- This models are constructed automatically from existing RNA sequence alignments or even from initially unaligned example sequence, using an iterative training procedure that is essentially an automatic implantation of comparative sequence analysis and an algorithm that we believe in the first optimal algorithm for RNA secondary structure prediction on pairwise covariations n multiple alignments.

  We test these algorithms using data taken from a trusted alignment of this t RNA sequences(12) and on genomic sequence data from the C elegans genome sequencing project.

  We find that an automatically constructed t RNA covariance models a significantly more sensitive for databse searching that even the best custom built t RNA searching program.

- Our method produce t RNA alignments of higher accuracy than other automatic methods and they invariably predict the correct consensus cloverleaf t RNA secondary structure when given unaligned. Example t RNA sequence.

Application of RNA structure modeling

•Knowing the shape of a biomolecule is invaluable in drug design and understanding disease mechanisms

•Current physical methods (X-Ray, NMR) are too expensive and time-consuming

•Predict shape from sequence of bases

UNIT 3

Machine learning-:

-machine learning is subfield of computer science and artificial intelligence that deals with the construction and study of systems that can learn from data ,rather than follow only specific explicitly programs instruction.

-besides computer science and artificial intelligence it has strong statistics and optimization which deliver both methods and theory to the field.

Examples= application includes spam filtering, optical character recognition,search engines and computer viruses.

-machine learning, data mining,pattern recognition are sometimes conflated. Machine learning task can be of several forms –

1-supervised learning

2-unsupervised learning

1-SUPERVISED -: In supervised, the computer is presented with example inputs and their desired outputs given by a 'teacher' inputs to outputs.

Spam filtering is an ex of supervised learning.

2-UNSUPERVISED-:In unsupervised learning , no labels are given to the algorithm ,leaving it on its own to groups of similar inputs (clustering),density estimates or projection of high dimensional data that can be visualized effectively.

-unsupervised learning can be a goal in itself (discovering hidden patterns in data).

Example= topic modelling is an ex of unsupervised learning where a program is given a lot of language documents and is tasked to find which document covers similar topics.

3.In REINFORCEMENT LEARNING,a computer program interacts with a dynamic environment in which it perform a certain goal such as driving a vehicle without a teacher explicitly telling it whether it has come close to its goal or not.

Definition of machine learning

In 1959, Arthus Samuel defined machine learning as a field of that gives computers the ability to learn without the explicitly program.

DECISION TREE INDUCTION-:

Decision tree is a simple and widely used classification technique.

Suppose a new species is discovered by scientist how can be tell whether it is a mammal or a non-mammal?

One approach is to pose a series of questions about the characteristics of the species. The first question we may ask is whether the species is cold or warm blooded. If it is not a mammal .Otherwise it is either a bird or

mammal. In a latter case we need to ask a follow up question: - to the females of the species give birth to their young?

Those that give birth are definitely mammals while those that do not are likely to be non-mammals (with the exception of lay egging mammals such as the spiny and eater).

The previous example demonstrate how we can solve a classification problem by asking a series of questions about the attributes of the test records. Each time we receive an answer, a follow up question is asked until we reach conclusion about the class label of the record. The series of question and their possible answer can be organized in the form of a decision tree, which hierarchical structure consisting of nodes and directed edges.

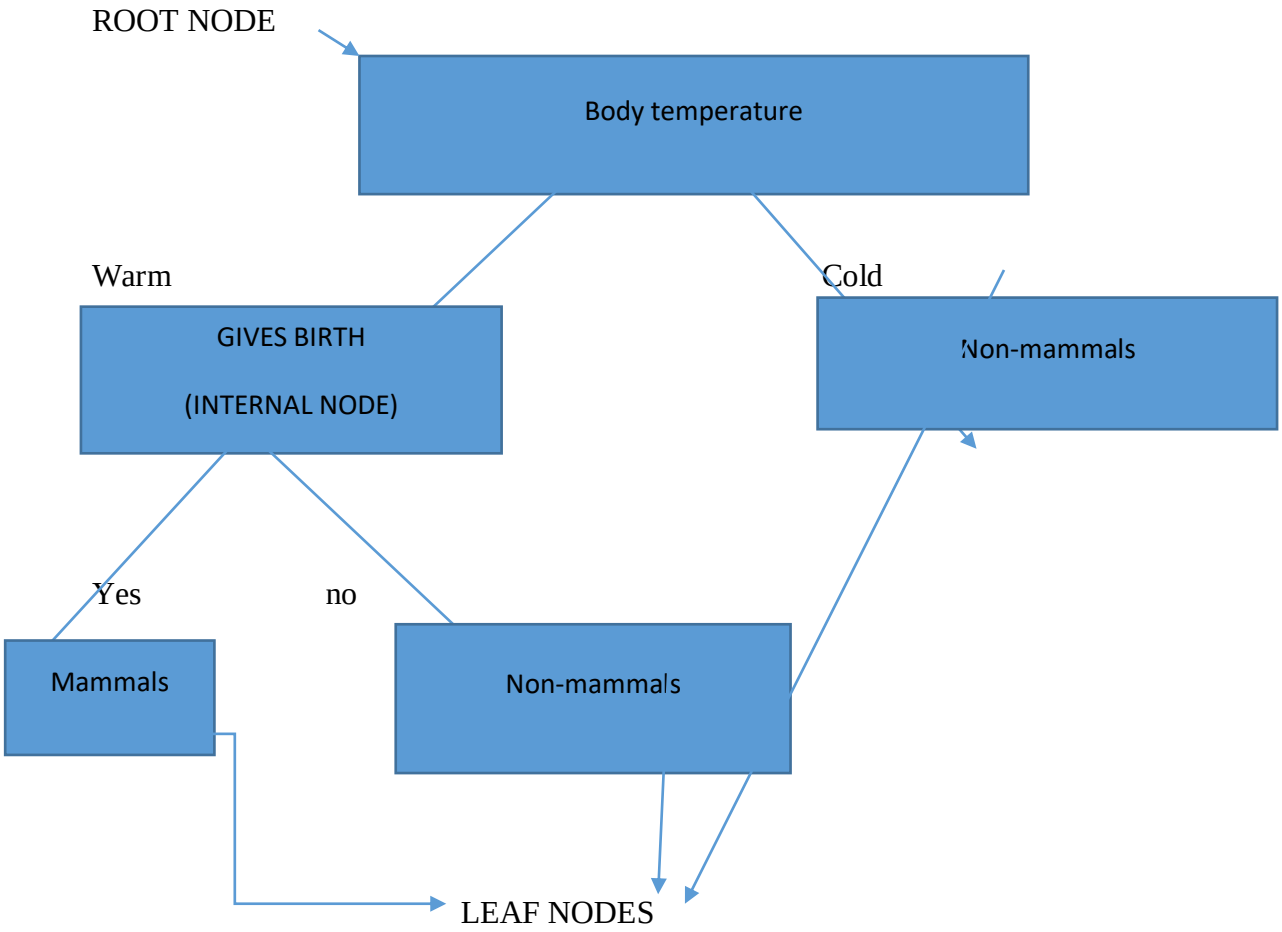


FIG-:A decision tree for classification of mammals.

The tree has three types of nodes

1-ROOT NODE

2-INTERNAL NODE

3-LEAF NODE OR TERMINAL NODE

-A root node that has no incoming edges and zero or more outgoing edges.

-INTERNAL NODE: each of which have exactly one incoming edge and two or more outgoing edges.

-LEAF NODE: each of which have exactly one incoming edge and no outgoing edges.

ARTIFICIAL NEURAL NETWORK

-Artificial neural network is inspired from the natural neural network of human nervous system.

-The inventor of the 1st neuro computer is Dr. ROBERT HETCH-NIELSON defines a neural network as –

"A computing system made up of number of highly processing elements, which process information by their dynamic state response to internal inputs.

Basic structure of artificial neural network-:

-The idea of ANN is based on the belief that the working of human brain by making connections, can be initiated using silicones and wires as living neurons and dendrites.

-The human brain is composed of 86 billion nerve cells called neurons. They are connected to other thousands cells by axons .Stimuli from external environment or inputs from sensory organs are accepted by dendrites. These inputs create electric impulses which quickly travel through the neural network.

-A neuron can then send the message to other neurons to handle the issue.

-ANN are composed of multiple nodes which initiate the biological neurons of human brain. The neurons are connected by links and they interact with each other. The nodes can take input data and perform simple operations on the data. The result of these operations is passed through other neurons. The outputs at each node is called its ACTIVATION OR NODE VALUE.

FIG: Simple ANN

Types of artificial neural networks:-

1-feed forward ANN.

2-feedback ANN.

1-FEED FORWARD ANN-:

-In this ANN the information flow is unidirectional. A unit sense information to other unit from which it does not receive any information. There are no feedback loops.

-They are used in pattern generation /recognition /classification.

-They have fixed inputs and outputs.

2-FEED BACKWARD ANN-:

Here feedback loops are formed.

Working of ANN-:

-In the topology diagrams shown each arrow represents a components a connection between two neurons and indicates the pathway for the flow of information. Each connection has a weight, an integer number that controls the signals between 2 neurons.

-If the network generates a good or desired output there is no need to adjust the weights. However if the network generates a poor or undesired output or an error than the system utters the weight in order to remove subsequent results.

MACHINE LEARNING IN ANN

ANN are capable of learning and they used to be trained. There are several learning strategies-:

1-supervised learning

2-unsupervised learning

3-reinforcement learning

Artificial intelligence issue

-Artificial intelligence is developing with such an incredible speed, sometimes it seems to be maximal. This is an opinion among researchers and developers that artificial intelligence would grow so inversely strong that it would be difficult for humans to control.

-Humans develop artificial intelligence systems by introducing into them every possible intelligence. They would, for which the humans themselves now seems to be threatened.

1- THREAT TO PRIVACY.
2- THREAT TO HUMAN DIGNITY.
3- THREAT TO SAFETY.

1-THREAT TO PRIVACY- An artificial intelligence program that recognizes speech and understand natural language is theoretically capable of understanding each conversation on emails and telephones.

HIDDEN MARKOV MODEL

-Hidden Markov model is a statistical markov model in which the system being modeled is assumed to be markov model with unobserved (hidden) states.
-The hidden markov model can be represented as the simplest bayesion network.
The mathematics behind the HMM work develop by L.E Baum and coworkers.
-In simpler markov model, the states is directly visible to observer and therefore the states transition probability are the only parameters while in HMM the state is not directly visible but the output dependent on the states is visible.
-HMM are specially known for their application in reinforcement learning and temporal patterns such as speech and writing gesture recognition and in bioinformatics.
A- Markov model
Example- 1- talk about the weather

1- Assume there are 3 weather (sunny, rainy & foggy).
3-Weather predictions is about the would the weather tomorrow. (based on the observations in the past)

4-Weather at day 'n' is

[Qn epsilon(sunny, rainy,foggy)]

Where qn = depends on the known weather of past days(qn-1,qn-2,-----).

5- We want to find that
P(qn, qn-1,qn-2,----- q1) - means given the past weather what is the probability of any possible weather of today.

B- Hidden Markov model

1- Suppose you locked in a room for several days.
2- We will try to predict the weather outside.
3- The only piece of evidence we have this wheather the person who comes into a room bringing our daily need is carrying an umbrella or not.
4- Finding the probability of an certain weather qn epsilon(sunny, rainy, foggy) based on the observations x1.
5- Using bayes rules-
[P(qi/xi)=((P (xi/qi) * P(qi))/ P(xi))]

6- For n days –[P(q1---qn/x1----xn)=((P(x1---xn/q1---qn) * P(q1---qn)) / P(x1---xn))].

Applications of HMM-:

1- Computational finance.
2- Speech recognition.
3- Gene determination.

4- Gene prediction.
5- Handwriting recognition.
6- Protein folding.
7- Sequence classification.
8- DNA motif discoverey.
9- Transportation forecasting.
10- Machine translation.
11- Single molecule kinetic analysis.

GENETIC ALGORITHM-:

Genetic Algorithms:-Genetic algorithms are examples of evolutionary computing methods and are optimizationtype algorithms. Given a population of potential problem solutions (individuals), evolutionary computing expands this population with new and potentially better solutions. They are the basis for evolutionary computing algorithms is biological evolution, where over time evolution produces the best or "fittest" individuals.

-In Data mining, genetic algorithms may be used for clustering, prediction, and even association rules.

-When using genetic algorithms to solve a problem, the first thing, and perhaps the most difficult task, that must be determined is how to model the problem as a set of individuals. In the real world, individuals may be identified by a complete encoding of the DNA structure.

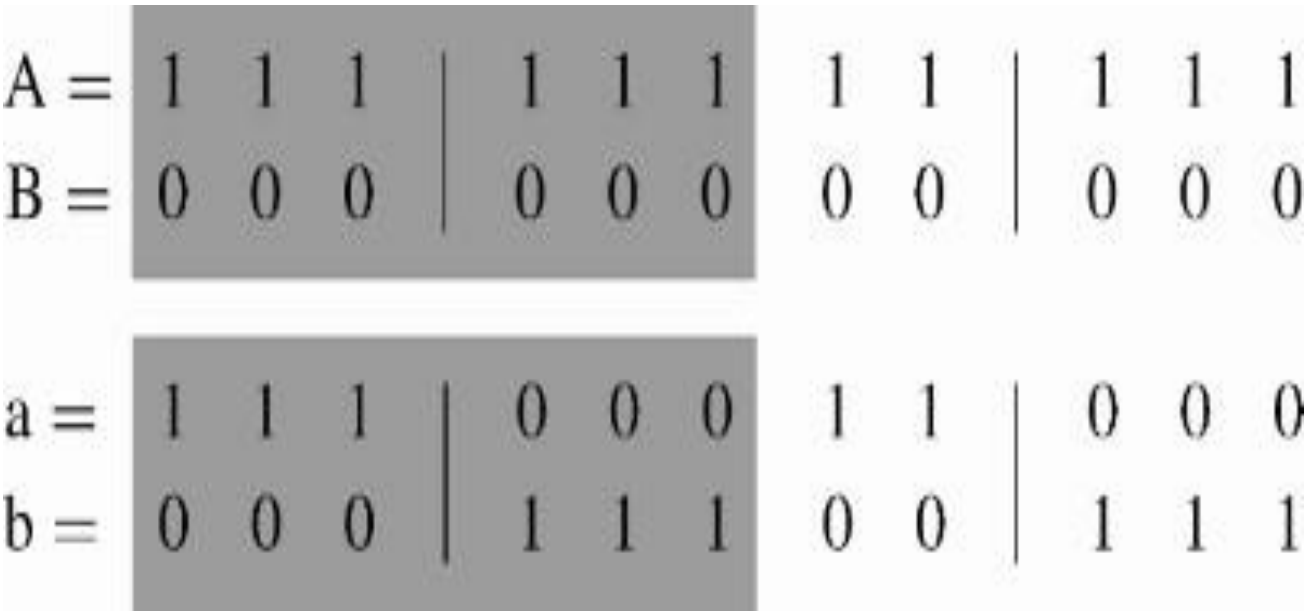
- An individual typically is viewed as an array or tuple of values. Based on the recombination (crossover) algorithms, the values are usually numeric and maybe binary strings.

These individuals are like a DNA encoding in the structure for each individual represents an encoding of the major features needed to model the problem. Each individual in the population is represented as a string of characters from the given alphabet.

Definition: Given an alphabet A, an **individual** or **chromosome** is a string I = I1, I2,…, In where Ij ϵ A. Each character in the string, Ij, is called a gene. The values that each character can have are called the alleles. A populations, P, is a set of individuals.

In genetic algorithms, reproduction is defined by precise algorithms that indicate how to combine the given set of individuals to produce new ones. These are called "crossover algorithms". For example: Given two individuals; parents from a population, the crossover technique generates new individuals (offspring or children) by switching subsequences of the string.

Genetic Algorithms



-As in nature, mutations sometimes appear, and these also may be present in genetic algorithms. The mutation operation randomly changes characters in the offspring and a very small probability of mutation is set to determine whether a character should change.

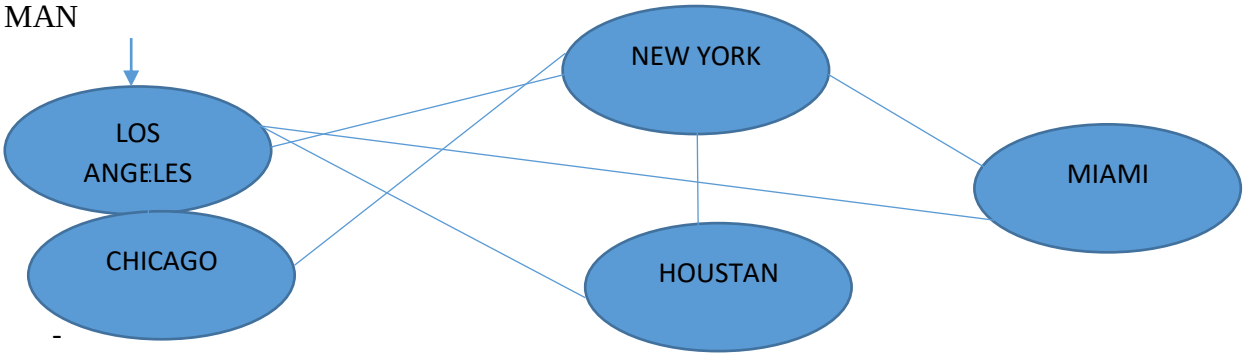
- Since genetic algorithms attempts to model nature, only the strong survive. When new individuals are created, a choice must be made about which individuals will survive. This may be the new individuals, the old ones, or more likely a combination of the two. It is the part of genetic algorithms that determines the best (or fittest) individuals to survive.
- To sum up all these information, Margaret Dunham defines a genetic algorithm (GA) as a computational model consisting of five part: Starting set of individuals.

Crossover technique.

- Mutation algorithm.
- Fitness function (survivor of the strongest)
- Algorithms that applies the crossover and mutation to P iteratively using the fitness function to determine the fitness function to determine the best individuals in P to keep.

Simulated annealing

What is simulated annealing ?
It is an algorithm for finding a good solution to an optimization problems.
What's an optimization problem?
It's a problem of finding the best solution from all the feasible solutions
Examples -: travelling salesman
MAN



-
- The salesman needs to minimize the number of miles he travels and itenarary is better if it is shorter .
- There are many feasible itenararies to choose from we are looking for the best one.
- NOTE-: simulated annealing solves this type of problems.

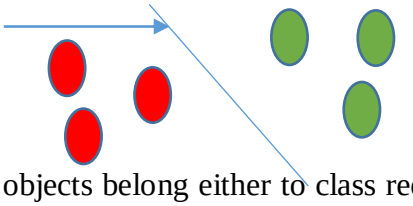
Why annealing?

-simulated annealing is inspired by a metal working process called annealing.
-it uses the equation that describes changes in a metals embodied, energy during the annealing process.
NOTE- Simulated annealing is generally considered a good choice for solving optimization problems.

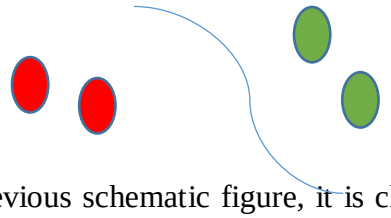
SUPPORT VECTOR MACHINE (SVM)-:

-
- SVM are based on the concept of decision plans that define decision boundaries. A decision plan is one that separate between a set of objects having different class memberships.
  Fig-: a schematic example is shown here
  decision plane

  red



- In this example, objects belong either to class red or green. The separating line defines the boundary on the right side of which all objects are green and to the left of which all objects are red. Any new objects falling to the right is labelled or classified as green and vice versa.
- The above is a classical example of linear classifier, i.e. a classifier that separates a set of objects into their respective groups (green and red in this case) with a line . most classification task are not so simple and often more complex structures are needed in order to make an optimal separation, i.e. correctly classify new objects on the basis of examples that are available .



- Compare to the previous schematic figure, it is clear that a full separation of the green and red objects could require a curve (which is more complex than a line).
- Classification task based on drawing separating lines to distinguish between objects of different class membership are known as hyper plain classification.
- SVM are particularly suited to handle such task.

NOTE- SVM is primarily a classifier method that performs classification task by constructing hyper plain in a multidimensional space that separates cases of different class levels.

Empirical researcher and empirical cycle-:

→ Empirical research is a research using empirical evidence.
→ It is a way of gaining knowledge by means of direct or indirect observations or experience.
→ Empirical evidence (the record of once direct observation or experience) can be analyze quantitatively and qualitatively.

Quantifying the evidencefor making sense of it in quantitative form,the researcher can answer empirical questions which should be clearly defined and answerable with evidence collected (usually all data). Many researchers can combine qualitative and quantitative form of analysis to better answer question which cannot to studied in laboratory settings and in education.

In some fields quantitative research may begin with a research question which is texted through experimental. Usually researchers has a certain theory regarding a topic under investigation .Based on

this theory statement or hypothesis will be proposed. From this hypothesis,prediction about specific events are derived. These predictions can then be texted with a suitable example experiment depending on the outcome of the experiment,the theory on which the hypothesis and prediction were based will be supported or not or may need to be modified and then subjected to further texting.

Empirical cycle used in empirical research.

STEPS IN EMPIRICAL RESEARCH

1. PROBLEM STATEMENT, PURPOSES, BENEFITS, what exactly do I want to find out? What is a researchable problem? What are the obstacles in terms of knowledge, data availability, time, or resources? Do the benefits outweigh the costs?

2. THEORY, ASSUMPTIONS, BACKGROUND LITERATURE, what does the relevant literature in the field indicate about this problem? To which theory or conceptual framework can I link it? What are the criticisms of this approach, or how does it constrain the research process? What do I know for certain about this area? What is the history of this problem that others need to know?

3. VARIABLES AND HYPOTHESES, what will I take as given in the environment? Which are the independent and which are the dependent variables? Are there control variables? Is the hypothesis specific enough to be researchable yet still meaningful? How certain am I of the relationship(s) between variables?

4. OPERATIONAL DEFINITIONS AND MEASUREMENT, what is the level of aggregation? What is the unit of measurement? How will the research variables be measured? What degree of error in the findings is tolerable? Will other people agree with my choice of measurement operations?

5. RESEARCH DESIGN AND METHODOLOGY, what is my overall strategy for doing this research? Will this design permit me to answer the research question? What other possible causes of the relationship between the variables will be controlled for by this design? What are the threats to internal and external validity?

6. SAMPLING, how will I choose my sample of persons or events? Am I interested in representativeness? If so, of whom or what, and with what degree of accuracy or level of confidence?

7. INSTRUMENTATION, how will I get the data I need to test my hypothesis? What tools or devices will I use to make or record observations? Are valid and reliable instruments available, or must I construct my own?

8. DATA COLLECTION AND ETHICAL CONSIDERATIONS, are there multiple groups, time periods, instruments, or situations that will need to be coordinated as steps in the data collection process? Will interviewers, observers, or analysts need to be trained? What level of inter-rater reliability will I accept? Do multiple translations pose a potential problem? Can the data be collected and subjects' rights still preserved?

9. DATA ANALYSIS, what combinations of analytical and statistical process will be applied to the data? Which will allow me to accept or reject my hypotheses? Do the finding show numerical differences, and are those differences important?

10. CONCLUSIONS, INTERPRETATIONS, RECOMMENDATIONS, was my initial hypothesis supported? What if my findings are negative? What are the implications of my findings for the theory base, for the background assumptions, or relevant literature? What recommendations can I make for public policies or programs in this area? What suggestions can I make for further research on this topic?

Clustering:-

Clustering can be considered the most important unsupervised learning problem, so as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.

A definition of clustering could be "the process of organizing objects into groups whose members are similar in some way".

A cluster is therefore, a collection of object which are similar in between thenand are dissimilar to the objects belonging to other clusters.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters.

Let's understand this with an example. Suppose, you are the head of a rental store and wish to understand preferences of your costumers to scale up your business. Is it possible for you to look at details of each costumer and devise a unique business strategy for each one of them? Definitely not. But, what you can do is to cluster all of your costumers into say 10 groups based on their purchasing habits and use a separate strategy for costumers in each of these 10 groups. And this is what we call clustering.

Now, that we understand what is clustering. Let's take a look at the types of clustering.

Types of Clustering

Broadly speaking, clustering can be divided into two subgroups :

- **Hard Clustering:** In hard clustering, each data point either belongs to a cluster completely or not. For example, in the above example each customer is put into one group out of the 10 groups.

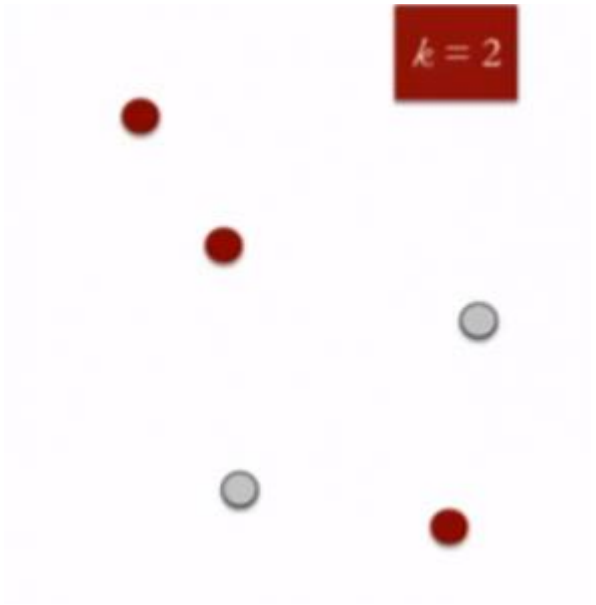
- **Soft Clustering**: In soft clustering, instead of putting each data point into a separate cluster, a probability or likelihood of that data point to be in those clusters is assigned. For example, from the above scenario each costumer is assigned a probability to be in either of 10 clusters of the retail store.

Now I will be taking you through two of the most popular clustering algorithms in detail – K Means clustering and Hierarchical clustering. Let's begin.

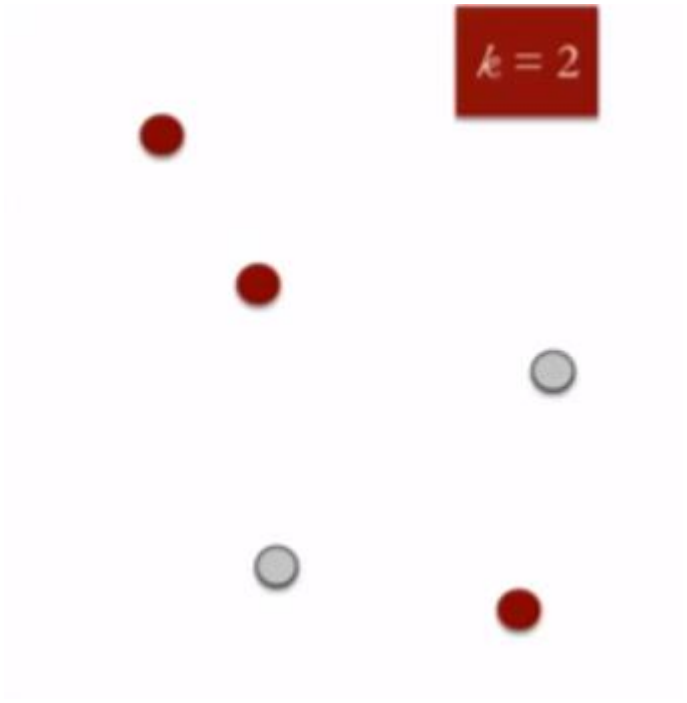
K Means Clustering

K means is an iterative clustering algorithm that aims to find local maxima in each iteration. This algorithm works in these 5 steps :

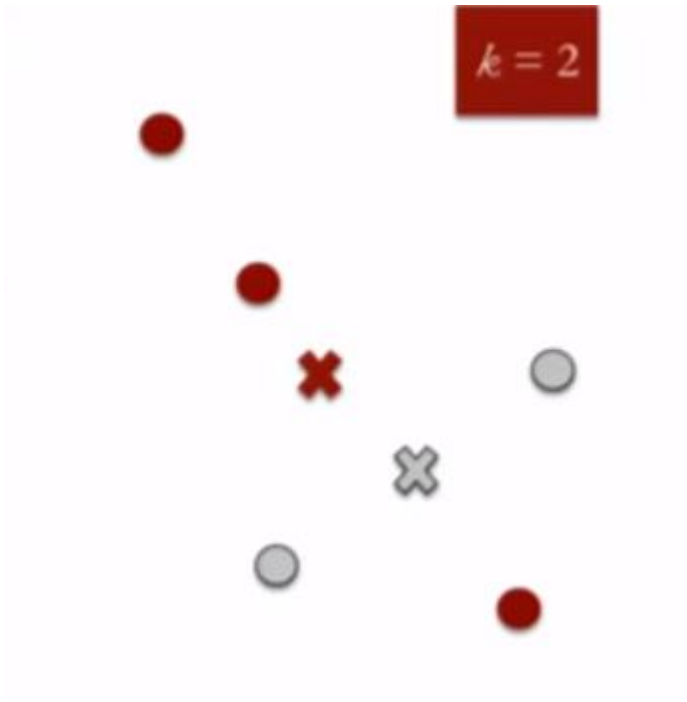
1. Specify the desired number of clusters K : Let us choose k=2 for these 5 data points in 2-D space.



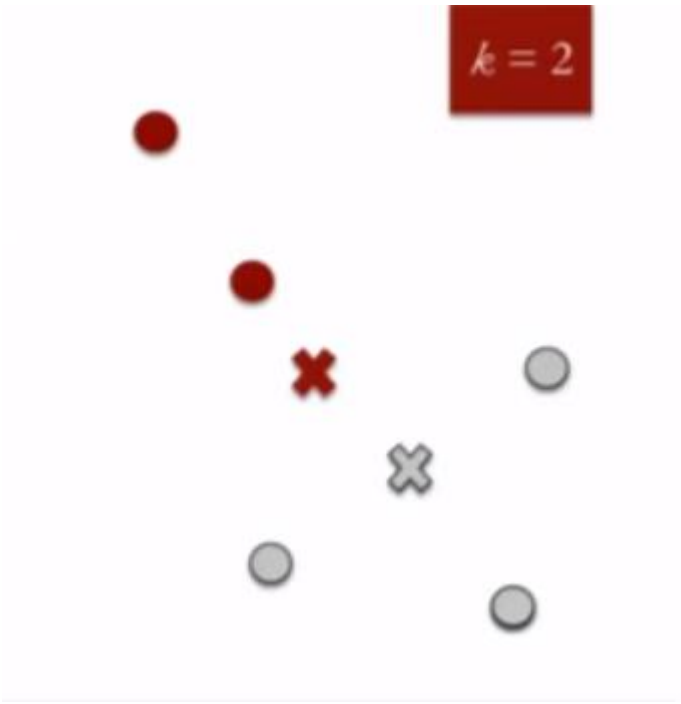
2. Randomly assign each data point to a cluster : Let's assign three points in cluster 1 shown using red color and two points in cluster 2 shown using grey color.



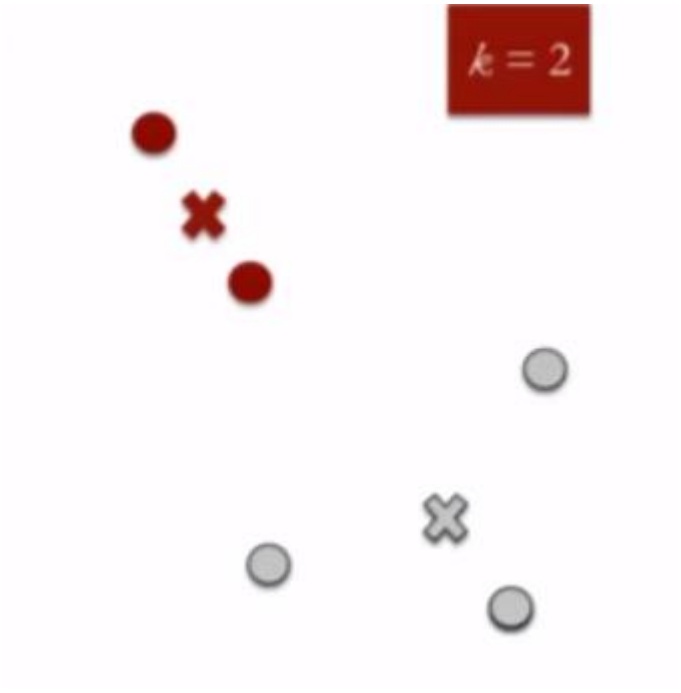
3. Compute cluster centroids : The centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.



4. Re-assign each point to the closest cluster centroid : Note that only the data point at the bottom is assigned to the red cluster even though its closer to the centroid of grey cluster. Thus, we assign that data point into grey cluster



5. Re-compute cluster centroids : Now, re-computing the centroids for both the clusters.

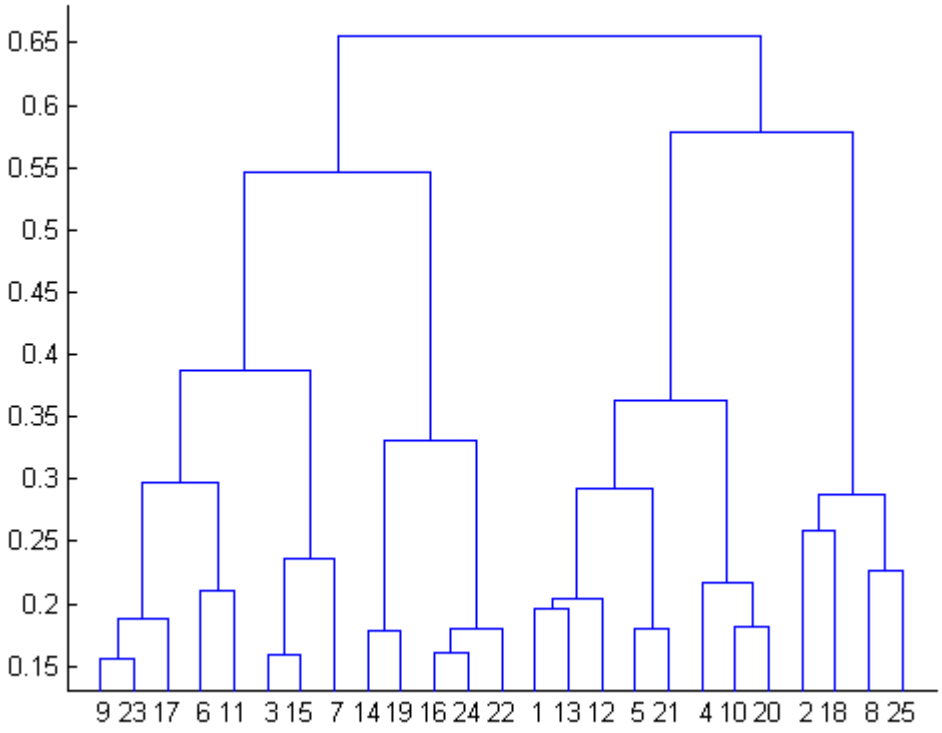


6. Repeat steps 4 and 5 until no improvements are possible : Similarly, we'll repeat the 4th and 5th steps until we'll reach global optima. When there will be no further switching of data points between two clusters for two successive repeats. It will mark the termination of the algorithm if not explicitly mentioned.

Hierarchical Clustering

Hierarchical clustering, as the name suggests is an algorithm that builds hierarchy of clusters. This algorithm starts with all the data points assigned to a cluster of their own. Then two nearest clusters are merged into the same cluster. In the end, this algorithm terminates when there is only a single cluster left.

The results of hierarchical clustering can be shown using dendrogram. The dendrogram can be interpreted as:



Applications of Clustering

Clustering has a large no. of applications spread across various domains. Some of the most popular applications of clustering are:

- Recommendation engines
- Market segmentation
- Social network analysis
- Search result grouping
- Medical imaging
- Image segmentation
- Anomaly detection

Evaluation of prediction methods: Parametric and Non- parametric tools:-

PARAMETRIC MACHINE LEARNING ALGORITHM:-

Assumptions can greatly simplify the learning process, but can also limit what can be learned. Algorithms that simplify the function to a known form are called parametric machine learning algorithms.

NOTE: A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called as parametric model. No matter how much data you throw at a parametric model, it won't change its mind about how many parameters it needs.

Examples: 1. Logistic regression
2- Linear discriminant analysis.
3- Perception
4- Naïve Bayes
5- Simple neural network

Benefits-

Simpler, speed and less data.

Limitations-

Constrained, limited complexity and poorfit.

NON-PARAMETRIC MACHINE LEARNING ALGORITHM:-

Algorithms that do not make strong assumptions about the form of the mapping functions are called non-parametric learning algorithms.
Examples: support vector machines.

Benefits-

Flexibility, power and performance.

Limitations- More data, slower and overfitting.

	Parametric	Non-parametric
Assumed distribution	Normal	Any
Assumed variance	Homogeneous	Any
Typical data	Ratio or Interval	Ordinal or Nominal
Data set relationships	Independent	Any
Usual central measure	Mean	Median
Benefits	Can draw more conclusions	Simplicity; Less affected by outliers
Tests		
Choosing	Choosing parametric test	Choosing a non-parametric test
Correlation test	Pearson	Spearman
Independent measures, 2 groups	Independent-measures t-test	Mann-Whitney test
Independent measures, >2 groups	One-way, independent-measures ANOVA	Kruskal-Wallis test
Repeated measures, 2 conditions	Matched-pair t-test	Wilcoxon test

The Relation between Statistics and Machine Learning

The machine learning practitioner has a tradition of algorithms and a pragmatic focus on results and model skill above other concerns such as model interpretability.

Statisticians work on much the same type of modeling problems under the names of applied statistics and statistical learning. Coming from a mathematical background, they have more of a focus on the behavior of models and explainability of predictions.

The very close relationship between the two approaches to the same problem means that both fields have a lot to learn from each other. The statisticians need to consider algorithmic methods was called out in the classic “two cultures” paper. Machine learning practitioners must also take heed, keep an open mind, and learn both the terminology and relevant methods from applied statistics.

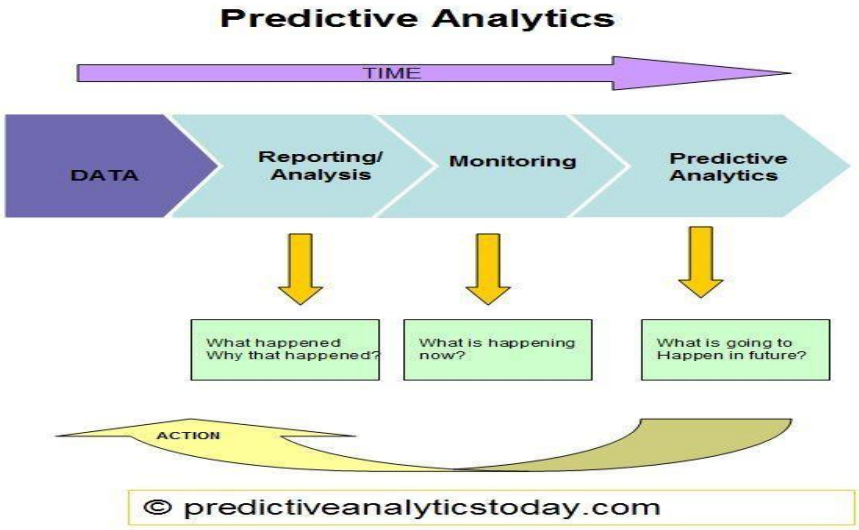
Predictive Analytics & Predictive Modelling

What is Predictive Modelling

➨ **Predictive analytics** is the branch of the advanced **analytics** which is used to make predictions about unknown future events. **Predictive analytics** uses many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions about future.

➨ **Predictive modeling** is a process used in **predictive** analytics to create a statistical **model** of future behavior. **Predictive** analytics is the area of data mining concerned with forecasting probabilities and trends.

Predictive Analytics Process



Business process and features on Predictive Modelling

 Business process on Predicting modelling

❖ Creating the model

❖ Testing the model

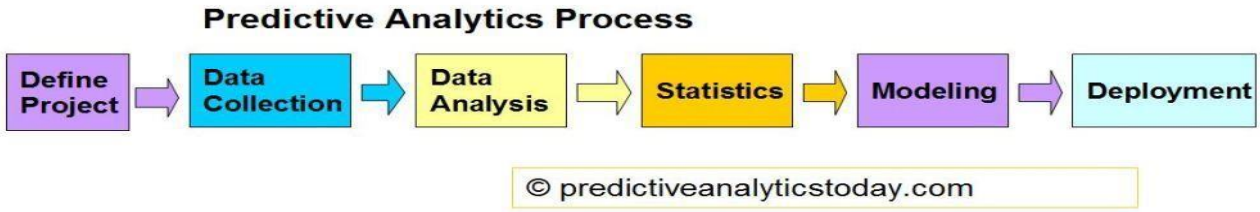
- Validating the model
- Evaluating the model

Features in Predicting modelling

- Data analysis and manipulation
- Visualization
- Statistics
- Hypothesis testing

How the model work

In predictive modeling, data is collected for the relevant predictors, a statistical model is formulated, predictions are made and the model is validated (or revised) as additional data becomes available. The model may employ a simple linear equation or a complex neural network, mapped out by sophisticated software.



How the model work(cont.)

Here you will learn what a predictive model is, and how, by actively guiding marketing campaigns, it constitutes a key form of business intelligence. we'll take a look inside to see how a model works-

1. **Predictors Rank Your Customers to Guide Your Marketing**
2. **Combined Predictors Means Smarter Rankings**
3. **The Computer Makes Your Model from Your Customer Data**
4. **A Simple Curve Shows How Well Your Model Works**
5. **Conclusions**

Why Predictive Modelling ?

Nearly every business in competitive markets will eventually need to do predictive modeling to remain ahead of the curve. Predicting Modeling (also known as Predictive Analytics) is the process of automatically detecting patterns in data, then using those patterns to foretell some event. Predictive models are commonly built to predict:

- **Customer Relationship Management**
- **the chance a prospect will respond to an ad**
- **Mail recipients likely to buy**
- **when a customer is likely to churn**
- **if a person is likely to get sick**
- **Portfolio or Product Prediction**
- **Risk Management & Pricing**

Some Predictive Models

Ideally, these techniques are widely used:

- Linear regression
- Logistic regression
- Regression with regularization
- Neural networks
- Support vector machines
- Naive Bayes models
- K-nearest-neighbors classification

- Decision trees
- Ensembles of trees
- Gradient boosting

Applications of Predictive Modelling

- Analytical customer relationship management (CRM)
- Health Care
- Collection Analytics
- Cross-cell
- Fraud detection
- Risk management

 Industry Applications

Predictive modelling are used in insurance, banking, marketing, financial services, telecommunications, retail, travel, healthcare, oil & gas and other industries.

Predictive Models in Retail industry

 Campaign Response Model – this model predicts the likelihood that a customer responds to a specific campaign by purchasing a products solicited in the campaign. The model also predicts the amount of the purchase given response.

- Regression models
- Customer Segmentation
- Cross-Sell and Upsell
- New Product Recommendation
- Customer Retention/Loyalty/Churn
- Inventory Management

Predictive Models in Telecom industry

- Campaign analytics

- Churn modeling
- Cross-selling and up-selling
- Customer lifetime value analytics
- Customer segmentation
- Fraud analytics
- Marketing spend optimization
- Network optimization
- Price optimization
- Sales territory optimization

Predictive Analytics Software

- SAS Analytics
- R
- STATISTICA
- IBM Predictive Analytics
- MATLAB
- Minitab

UNIT-4

Basic concept of Force field in molecular modeling

(Potential energy calculation)

A force field (a special case of energy functions or interatomic potentials; not to be confused with force field in classical physics) refers to the functional form and parameter sets used to calculate the potential energy of a system of atoms or coarse-grained particles in molecular mechanics and molecular dynamics simulations. The parameters of the energy functions may be derived from experiments in physics or chemistry, calculations in quantum mechanics, or both.

(A) Functional form

The basic functional form of potential energy in molecular mechanics includes bonded terms for interactions of atoms that are linked by covalent bonds, and nonbonded (also termed noncovalent) terms that describe the long-range electrostatic and van der Waals forces. The specific decomposition of the terms depends on the force field, but a general form for the total energy in an additive force field can be written as , E total = E bonded + E nonbonded where the components of the covalent and noncovalent contributions are given by the following summations:

E bonded = E bond + E angle + E dihedral

E nonbonded = E electrostatics + E vanderwaals

(B) Bond stretching

As it is rare for bonds to deviate significantly from their reference values the Morse potential is seldom employed for molecular mechanics due to it not being efficient to compute. The most simplistic approaches utilize a Hooke's law formula:

$v(l) = k/2 (l - l_0)^2$

Where k is the force constant, l is the bond length and l_0 is the value for the bond length when all other terms in the force field are set to 0. The term l_0 is often referred to as the equilibrium bond length which may cause confusion. The equilibrium bond length would be the value adopted in a minimum energy structure with all other terms contributing.

(C) Parametrization

In addition to the functional form of the potentials, force fields define a set of parameters for different types of atoms, chemical bonds, dihedral angles and so on. The parameter sets are usually empirical. A force field would

include distinct parameters for an oxygen atom in a carbonyl functional group and in a hydroxyl group. The typical parameter set includes values for atomic mass, van der Waals radius, and partial charge for individual atoms, and equilibrium values of bond lengths, bond angles, and dihedral angles for pairs, triplets, and quadruplets of bonded atoms, and values corresponding to the effective spring constant for each potential.

(D) Popular force fields

- AMBER, CHARMM, and GROMOS have been developed mainly for molecular dynamics of macromolecules.
- Assisted Model Building and Energy Refinement (AMBER)
- Chemistry at HARvard Molecular Mechanics (CHARMM)
- GROningenMOlecular Simulation (GROMOS)

INTRODUCTION TO SIMULATION

Definition :-

- A simulation is the imitation of the operation of real-world process or system over time.
- Generation of artificial history and observation of that observation history
- A model construct a conceptual framework that describes a system
- The behavior of a system that evolves over time is studied by developing a simulation model.
- The model takes a set of expressed assumptions:
- Mathematical, logical and Symbolic relationship between the entities

Goal of modelling

- A model can be used to investigate a wide verity of "what if" questions about real-world system.
- Potential changes to the system can be simulated and predicate their impact on the system.
- Find adequate parameters before implementation
- So simulation can be used as
- Analysis tool for predicating the effect of changes
- Design tool to predicate the performance of new system

- It is better to do simulation before Implementation.

How a model can be develop?

- Mathematical Methods
- Probability theory, algebraic method ,…
- Their results are accurate
- They have a few Number of parameters
- It is impossible for complex systems
- Numerical computer-based simulation
- It is simple
- It is useful for complex system

When simulation is the appropriate tool:-

- Simulation enable the study of internal interaction of a subsystem with complex system
- Informational, organizational and environmental changes can be simulated and find their effects
- A simulation model help us to gain knowledge about improvement of system
- Finding important input parameters with changing simulation inputs
- Simulation can be used with new design and policies before implementation

- Simulating different capabilities for a machine can help determine the requirement
- Simulation models designed for training make learning possible without the cost disruption
- A plan can be visualized with animated simulation
- The modern system (factory, wafer fabrication plant, service organization) is too complex that its internal interaction can be treated only by simulation.

When simulation is not appropriate :-

- When the problem can be solved by common sense.
- When the problem can be solved analytically.
- If it is easier to perform direct experiments.
- If cost exceed savings.
- If resource or time are not available.
- If system behavior is too complex.
- Like human behavior

<u>Advantages and disadvantages of simulation</u>

- In contrast to optimization models, simulation models are “run” rather than solved.
- Given as a set of inputs and model characteristics the model is run and the simulated behavior is observed

Advantages :-

- New policies, operating procedures, information flows and son on can be explored without disrupting ongoing operation of the real system.

- New hardware designs, physical layouts, transportation systems and … can be tested without committing resources for their acquisition.

- Time can be compressed or expanded to allow for a speed-up or slow-down of the phenomenon(clock is self-control).

- Insight can be obtained about interaction of variables and important variables to the performance.

- Bottleneck analysis can be performed to discover where work in process, the system is delayed.

- A simulation study can help in understanding how the system operates.

- "What if" questions can be answered.

Disadvantages:-

- Model building requires special training.

- Vendors of simulation software have been actively developing packages that contain models that only need input (templates).

- Simulation results can be difficult to interpret.

- Simulation modeling and analysis can be time consuming and expensive.

- Many simulation software have output-analysis.

Areas :-

- Manufacturing Applications

- Semiconductor Manufacturing

- Construction Engineering and project management

- Military application

- Logistics, Supply chain and distribution application

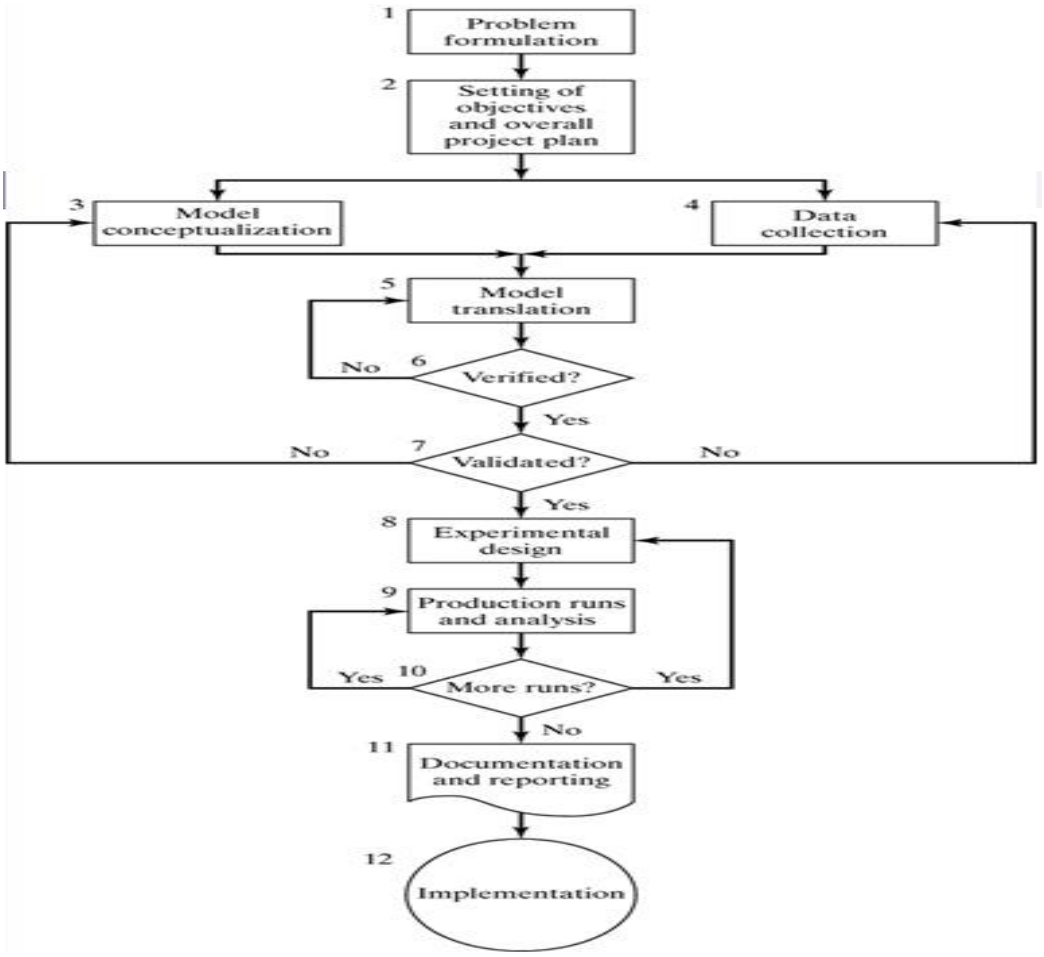
- Transportation modes and Traffic

- Business Process Simulation

- Health Care

- Automated Material Handling System (AMHS)
- Test beds for functional testing of control-system software
- Risk analysis
- Insurance, portfolio,...
- Computer Simulation
- CPU, Memory,…
- Network simulation
- Internet backbone, LAN (Switch/Router), Wireless, PSTN (call center),...

Steps in simulation study:-



Types of Simulation

Three types→Continuous, Discrete and Hybrid

- Continuous simulation is a simulation based on continuous time, rather than discrete time steps, using numerical integration of differential equations.
  It is notable as one of the first uses ever put to computers, dating back to the Eniac in 1946. Continuous simulation allows prediction of

- rocket trajectories
- hydrogen bomb dynamics
- electric circuit simulation
- robotic

- A discrete-event simulation (DES) models the operation of a system as a (discrete) sequence of events in time. Each event occurs at a particular instant in time and marks a change of state in the system.[1] Between consecutive events, no change in the system is assumed to occur; thus the simulation time can directly jump to the occurrence time of the next event, which is called next-event time progression.
  Example

  A common exercise in learning how to build discrete-event simulations is to model a queue, such as customers arriving at a bank to be served by a teller. In this example, the system entities are Customer-queue and Tellers. The system events are Customer-Arrival and Customer-Departure. (The event of Teller-Begins-Service can be part of the logic of the arrival and departure events.) The system states, which are changed by these events, are Number-of-Customers-in-the-Queue (an integer from 0 to n) and Teller-Status (busy or idle). The random variables that need to be characterized to model this system stochastically are Customer-Interarrival-Time and Teller-Service-Time. An agent-based framework for performance modeling of an optimistic parallel discrete event simulator is another example for a discrete event simulation

- Hybrid Simulation (sometime Combined Simulation) corresponds to a mix between Continuous and Discrete Event Simulation and results in integrating numerically the differential equations between two sequential events to reduce the number of discontinuities.

ALTERNATIVE ANIMAL USE

- In 1959 Russell and Burch introduced the concepts of replacement, reduction and refinement as a starting point for the humane treatment of laboratory animals.

The 3R's are:

- Replacement
- To replace the use of animals by lower organisms, or non-biological methods.

Refinement

To refine scientific procedures so that any suffering, discomfort or pain endured by the animals is minimal, consistent with the scientific objective.

Reduction

To reduce the number of animals used to the minimum consistent with statistical and scientific validity.

COMPUTER SIMULATION:-

- A computer model is the algorithms and equations used to capture the behavior of the system being modeled.
- By contrast, computer simulation is the actual running of the program that contains these equations or algorithms.
- In-silico -"performed on computer or via computer simulation"
- The phrase was coined in 1989 as an allusion to the Latin phrases in-vivo, in-vitro, andin-situ

ADVANTAGES

- Reduction
- Refinement
- Replacement
- Translation – “results of animal experimentation to human”

<u>COMPUTER SIMULATION TECHNIQUES IN VARIOUS FIELDS:-</u>

IN EDUCATION:-

->COMPUTER MODELS USED IN HUMAN PHYSIOLOGICAL STUDIES INCLUDES:-

- a comprehensive physiological model
- pH and carbon dioxide regulation
- pulsatile hemodynamics in the aorta
- determinants of cardiac output
- effects of medically important drugs on the circulatory system
- simulation of the digestion of a meal
- responses of organisms to exposure to high and low temperatures
- influence of hormones on muscle cells
- renal excretory response to volume and osmolarity changes

LIMITATIONS

- Biological complexities
- Missed experiences
- Biological variability
- Publication of results
- Student attitudes

IN BIOMEDICAL RESEARCH:-

A broad area of science that involves the investigation of the biological process and the causes of disease, through careful experimentation, observation, laboratory work, analysis, and testing

SOME EXAMPLES OF COMPUTER SIMULATIONIN BIOMEDICAL RESEARCH

- Kidney function –transport of electrolytes and water in and out of kidney
- Cardiac function –enzyme metabolism of cardiac muscle, cardiac pressure flow relationships, etc.
- Lung function –respiratory mechanics
- Sensory physiology –peripheral auditory system and single auditory nerve fibre transmission of vibrations
- Neurophysiology –impulse propagation along myelinated acxons
- Developmental biology –shape changes in embryonic cells

IN BEHAVIOURAL RESEARCH

Study of the variables that impact the formation of habits

SOME EXAMPLES OFCOMPUTER SIMULATIONOF BEHAVIORAL PHENOMENA

- Spacing mechanisms and animal movements
- Learning, memory, and problem solving
- Sensation and perception
- Communication
- Body maintenance
- Reproduction and parental care

LIMITATIONS

- Lack of knowledge of all possible parameters
- No satisfactory model exists
- Development of computer simulation depends on the use of animals in biomedical research

COMPUTER SIMULATIONIN OTHER FIELDS:-

DRUG DISCOVERY WITH VIRTUAL SCREENING

for example, using the protein docking algorithm EADock, researchers found potential inhibitors to an enzyme associated with cancer activityin silico

CELL MODELS

Efforts have been made to establish computer models of cellular behavior

for example, in 2007 researchers developed an in silico model of tuberculosis to aid in drug discovery, with the prime benefit of its being faster than real time simulated growth rates, allowing phenomena of interest to be observed in minutes rather than months

GENETICS

Digital genetic sequences obtained from DNA sequencing may be stored in sequence databases, be analyzed, be digitally altered and/or be used as templates for creating new actual DNA using artificial gene synthesis.

PROTEIN DESIGN

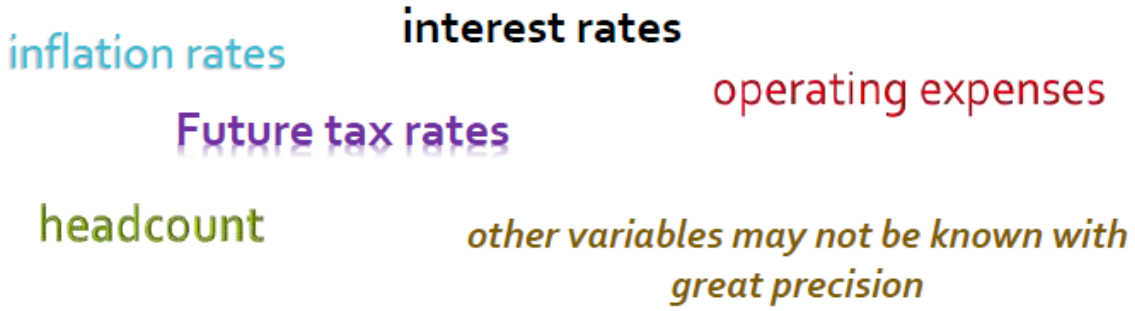
- example is RosettaDesign, a software package under development and free for academic use
- RosettaDesign can be used to identify sequences compatible with a given protein backbone

- Some of Rosetta design's successes include the design of a novel protein fold, redesign of an existing protein for greater stability, increased binding affinity between two proteins, and the design of novel enzymes

SENSITIVITY ANALYSIS-

Also called as what if analysis.

- Sensitivity analysis is the assessment of the impact for an output of a system by changing its inputs.
- example-In budgeting process there are always variables that are uncertain such as-



Sensitivity analysis answers the question of the above ques.

"if these variables deviate from expectations, what will the effect be (on the business, model, system, or whatever is being analyzed), and which variables are causing the largest deviations?"

TYPES OF SENSITIVE ANALYSIS

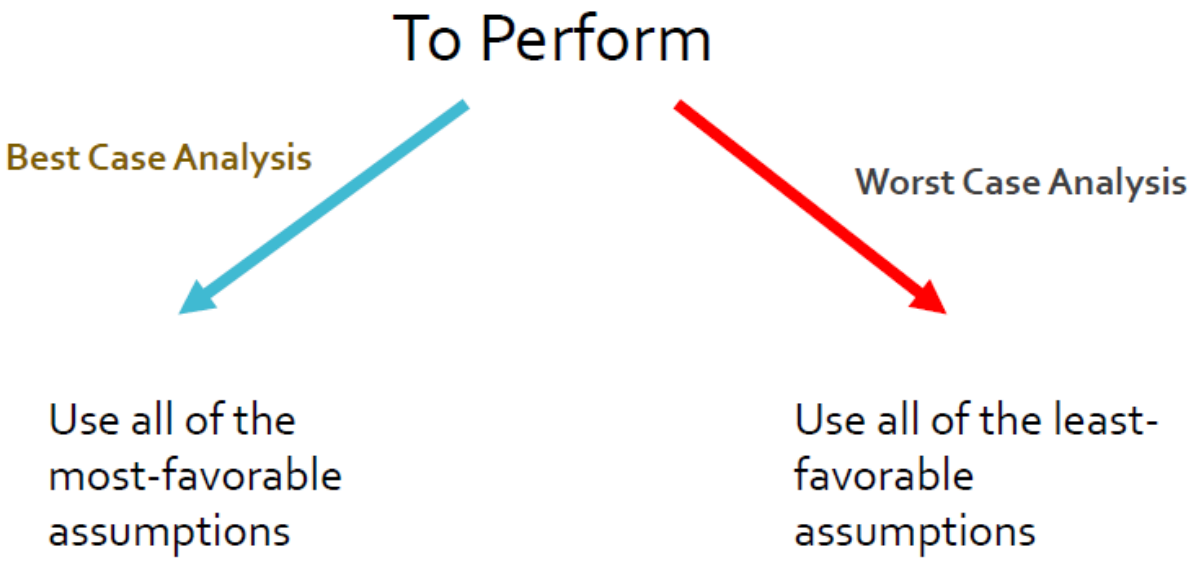
Partial Sensitivity Analysis

In a partial sensitivity analysis, you select one variable, change its value while holding the values of other variables constant.

Best-case and worst-case scenarios

Best- and worst-case scenarios establish the upper (best-case) and lower (worst-case) boundaries of a cost-benefit study's results.

This type of sensitivity analysis shows how a broad range of a program or policy's possible outcomes affect the bottom line.



- **Break-even analysis**

If you are unable to estimate a policy's most likely effects or cannot find comparable studies to help determine its best-case and worst-case scenarios, you can use break even analysis.

- **Monte Carlo analysis**

You can use Monte Carlo analysis to examine multiple variables simultaneously and simulate thousands of scenarios, resulting in a range of possible outcomes and the probabilities that they will occur.

ADVANTAGES

- Simplicity
- Directing Management Efforts
- Ease of being Automated
- As a quality Check.

DISADVANTAGES

- It does not provide clear cut results
- Not a solution in standalone form

UNIT5

Document clustering

Document clustering (or text clustering) is the application of cluster analysis to textual documents. It has applications in automatic document organization, topic extraction and fast information retrieval or filtering.

Document clustering involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document clustering is generally considered to be a centralized process. Examples of document clustering include web document clustering for search users.

The application of document clustering can be categorized to two types, online and offline. Online applications are usually constrained by efficiency problems when compared to offline applications.Text clustering may be used for different tasks, such as grouping similar documents (news, tweets, etc.) and the analysis of customer/employee feedback, discovering meaningful implicit subjects across all documents.

In general, there are two common algorithms. The first one is the hierarchical based algorithm, which includes single link, complete linkage, group average and Ward's method. By aggregating or dividing, documents can be clustered into hierarchical structure, which is suitable for browsing. However, such an algorithm usually suffers from efficiency problems. The other algorithm is developed using the K-means algorithm and its variants. Generally hierarchical algorithms produce more in-depth information for detailed analyses, while algorithms based around variants of the K-means algorithm are more efficient and provide sufficient information for most purposes.

These algorithms can further be classified as hard or soft clustering algorithms. Hard clustering computes a hard assignment – each document is a member of exactly one cluster. The assignment of soft clustering algorithms is soft – a document's assignment is a distribution over all clusters. In a soft assignment, a document has fractional membership in several clusters. Dimensionality reduction methods can be considered a subtype of soft clustering; for documents, these include latent semantic indexing (truncated singular value decomposition on term histograms) and topic models.

Other algorithms involve graph based clustering, ontology supported clustering and order sensitive clustering.

Given a clustering, it can be beneficial to automatically derive human-readable labels for the clusters. Various methods exist for this purpose.

Clustering in search engines

A web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories.

Procedures

In practice, document clustering often takes the following steps:

1. **Tokenization**

Tokenization is the process of parsing text data into smaller units (tokens) such as words and phrases. Commonly used tokenization methods include Bag-ofwords model and N-gram model.

2. **Stemming and lemmatization**

Different tokens might carry out similar information (e.g. tokenization and tokenizing). And we can avoid calculating similar information repeatedly by reducing all tokens to its base form using various stemming and lemmatization dictionaries. 3. Removing stop words and punctuation

Some tokens are less important than others. For instance, common words such as "the" might not be very helpful for revealing the essential characteristics of a text. So usually it is a good idea to eliminate stop words and punctuation marks before doing further analysis.

4. **Computing term frequencies or tf-idf**

After pre-processing the text data, we can then proceed to generate features. For document clustering, one of the most common ways to generate features for a document is to calculate the term frequencies of all its tokens. Although not perfect, these frequencies can usually provide some clues about the topic of the document. And sometimes it is also useful to weight the term frequencies by the inverse document frequencies. See tf-idf for detailed discussions.

5. **Clustering**

We can then cluster different documents based on the features we have generated. See the algorithm section in cluster analysis for different types of clustering methods.

6. **Evaluation and visualization**

Finally, the clustering models can be assessed by various metrics. And it is sometimes helpful to visualize the results by plotting the clusters into low (two) dimensional space. See multidimensional scaling as a possible approach.

Clustering v. Classifying

Clustering algorithms in computational text analysis groups documents into grouping a set of text what are called subsets or clusters where the algorithm's goal is to create internally coherent clusters that are distinct from one another. Classification on the other hand, is a form of supervised learning where the features of the documents are used to predict the "type" of documents.

Information retrieval system

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on fulltext or other content based indexing.

Information retrieval is the science of searching for information in a document, searching for documents themselves, and also searching for metadata that describe data, and for databases of texts, images or sounds. Automated information retrieval systems are used to reduce what has been called information overload. An IR systems is a software that provide access to books, journals and other documents, stores them and manages the document. Web search engines are the most visible IR applications.

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy. An object is an entity that is represented by information in a content collection or database. User queries are matched against the database information. However, as opposed to classical SQL queries of a database, in information retrieval the results returned may or may not match the query, so results are typically ranked. This ranking of results is a key difference of information retrieval searching compared to database searching.

Depending on the application the data objects may be, for example, text documents, images, audio,mind maps or videos. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata.

The two most important information retrieval systems in case of bioinformatics are as follows→

- ENTREZ in NCBI
- SRS in EMBL.

Concept of natural language

Definition of NLP-:

- Process information contained in natural language text.
- Also known as computational linguistic (CL),human language technology (HLT), natural language technology(NLE).

NLP FOR MACHINES-

- Analyze, understand and generate human languages just like humans do.

- Applying computational techniques to language domain

- To explain linguistic theories, to use the theories to build systems that can be of social use
- Started off as a branch of Artificial Intelligence
- Borrows from Linguistics, Psycholinguistics, Cognitive Science & Statistics.

WHY NLP is required ?

- A hallmark of human intelligence.
- Text is the largest repository of human knowledge and is growing quickly.
- Computer programmes that understood text or speech.

History of NLP-:

- In 1950, alan turing published an article "machine and intelligence" which advertised what is now called the turing test as a subfield of intelligence.
- Some beneficial and successful natural language systems were developed in the 1960s were SHRDLU, a natural language system working in restricted "blocks of words" with restricted vocabularies was written between 1964 to 1966.

Components of NLP-

1.NATURAL LANGUAGE UNDERSTANDING

-taking some spoken /typed sentence and working out what it means.
-mapping the given input in the natural language into a useful representation.
- different level of analysis required:
  Morphological analysis.
  Syntactic analysis.
  Semantic analysis.
  Discourse analysis.

2. **NATURAL LANGUAGE GENERATION**

- producing output in the natural language from some internal representation.
- different level of synthesis required:

Deep planning

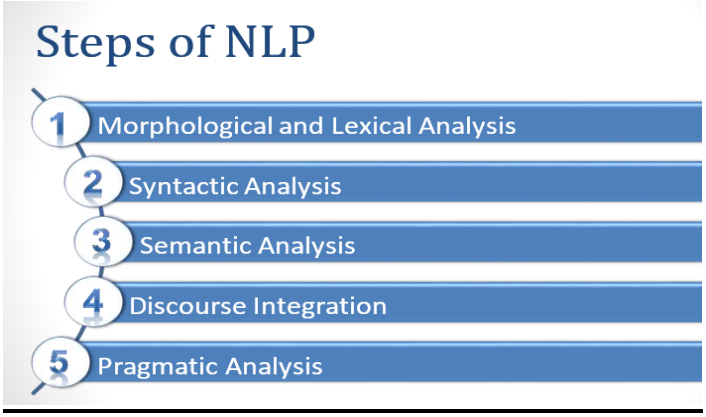
Syntactic generation

- NL understanding is much harder than NL generation. But , still both of them are hard.

Linguistics and language

- Linguistic is the science of language.
- Its study includes:
  . Sounds which refers to phonology.
  . Word formation refers to morphology
  . Sentence structure refers to syntax .
  . Meaning refers to semantics.
  . Understanding refers to pragmatics.

Steps of NLP



Morphological and lexical analysis

- The lexicon of a language is its vocabulary that includes its words and expressions.
- Morphology depicts analyzing, identifying and description of structure of words.
- Lexical analysis involves dividing a text into paragraphs , words and the sentences.

Syntactic analysis

- Syntax concerns the proper orderinf of words and its effect on meaning.
- This involves analysis of the words in a sentence to depict the grammatical structure of the sentence .
- The words are transformed into structure that shows how the words are related to each other. Ex- " the girl goes to the school". This would definitely be rejected by the English syntactic analyser.

Semantic analysis

-Semantic errors concerns the (literal)meaning of words, phrases and sentences.

- This abstract the dictionary meaning or the exact meaning from the context.
- The structures which are created by the syntactic analyzer are assigned meaning. Ex "colorless blue idea". This would be rejected by the analyzer as colorless blue do not make any sense together.

Discourse integration

- Sense of the context.

The meaning of any single sentence depends upon the sentences that precedes it and also invokes the meaning of the sentences that follow it. Ex- the word "it" in the sentence "she wanted it" depends upon the prior discourse context.

Pragmatic analysis

- Pragmatic analysis concerns the overallcommunicative and social context and its effect on interpretation.
- It means abstracting or deriving the purposeful use of the language in situations.
- Importantly those aspects of language which require world knowledge.
- The main focus is on what was said is reinterpreted on what it actually means. Ex- "close the window?". Should have been interpreted as a request rather than an order.

Natural language generation

- NLG is the process of constructing natural language outputs from non-linguistic inputs.
- NLG can be viewed as the reverse process of NL understanding.
- A NLG system may have two main parts:

  Discourse planner

  What will be generated. Which sentences.

  Surface realizer

  Realizes a sentence from its internal representation.

  Lexical selection

  Selecting the correct words describing the concepts.

Techniques and methods

- **Machine learning**

  . The learning procedures used during machine learning.

  . Automatically focuses on the most common cases.

  . Whereas when we write rules by hand it is often not correct at all.

  . Concerned on human errors.

- **Statistical interference**

  . Automatic learning procedures can make use of statistical interference algorithms.

  . Used to produce models that are robust (means strength) to unfamiliar input. Ex- containing words or structures that have not been seen before.

  . Making intelligent guesses.

- **Input database and training data**

  . Systems based on automatically learning the rules can be made more accurate simply by supplying more input data or source to it.

  . However, systems based on hand –written rules can only be made more accurate by increasing the complexity of the rules , which is a much more difficult task.

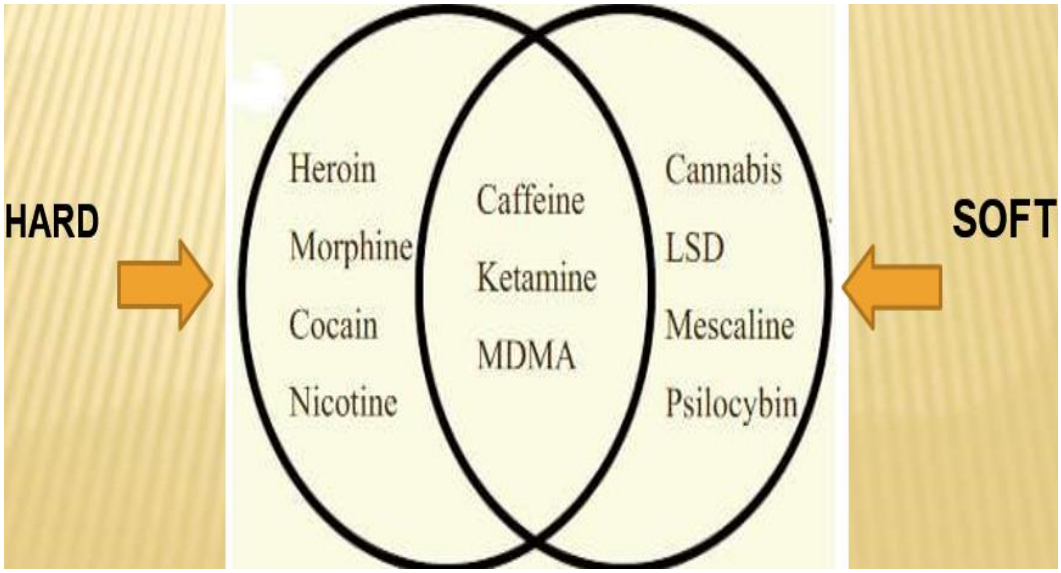
Concept of drug and drug designing

- A drug is any chemical you take that affects the way your bodyworks.
- Heroin, alcohol, ecstasy, caffeine and nicotine are all forms of drugs.
- A drug must be able to pass through your body and into your brain, allowing brain cells to be changed by interfering with the brains chemical signals called neurotransmitters.

Effects of illegal drugs-:

- Drugs make you less aware and alert making you feel carefree and can relieve pain.
- They can make you; sleep,have convulsions(fits) and even slip into a coma.

- For ex- heroin causes psychological and physical dependence which can lead onto comas and even worse death, as you can choke on your own vomit.
- <u>Hard and soft drugs</u>:
  Hard drugs- are physically addictive and may be easy to overdose on.
  Soft drugs – are not physically addictive.



Class: A, B & C

The different kinds of illegal drugs are split into 3classes: A, B & C. Each class carries a different level of punishment for possession and dealing

		Possession	*Dealing*
Class A	**Ecstasy, LSD, heroin, crack, magic fine or both**	**Up to 7 years in prison or an unlimited fine**	**Up to life in cocaine, mushrooms unlimited or both**
Class B	**Amphetamines, Methylphenidate (Ritalin),Pholcodine**	**Up to 5 years in prison an unlimited fine or both**	**Up to 14 years in prison or an unlimited fine or both**
Class C	**Cannabis, tranquilizers, painkillers, Gammahydroxybutyrate**	**Up to 2 years in prison or an unlimited fine or**	**Up to 14 years some unlimited fine**

Drug users-:

- If someone is using drugs, you might notice changes in how the person looks or acts. Here are some of the signs …
- Become moody, negative, cranky or worried all the time ask to be left alone a lot.
- Have trouble concentration.
- Loose or gain weight.
- Lose interest in hobbies.
- Change friends.
- Get in flights.
- Depression.

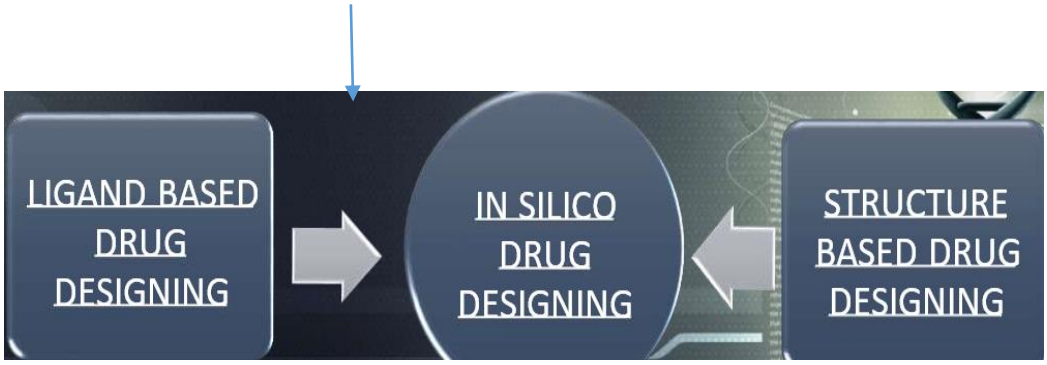
Insilico drug designing-

- Drug designing is the inventive method of finding new medications based on the knowledge of a biological target.
- Selected or designed molecule should be :
  Organic small molecule
  Complementary in shape.
  Oppositely charged to the biomolecule target.
- This molecule will –
  Interact with.

Bind to the target.

Activates or inhibits the function of a biomolecule such as a protein.

- In Silico is an expression used to mean “performed on computer or via computer simulation”.
- In Silico drug designing is defined as the identification of the drug target molecule by employing bioinformatics tools.
- TYPES OF INSILICO DRUG DESIGNING:



-

LIGAND BASED DRUG DESIGNING:

- It relies on knowledge other molecules that bind to the biological target of interest .
- used to derive a pharmacophore.

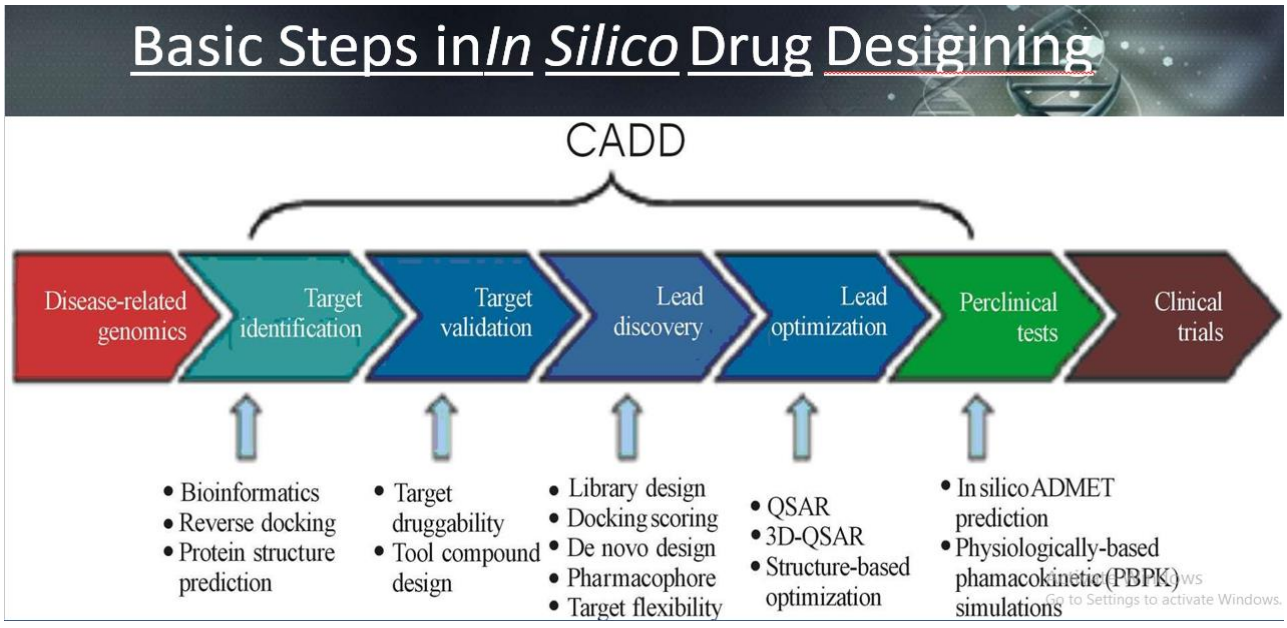
STRUCTURE BASED DRUG DESIGNING:

- It relies on the knowledge of the 3-D structure of a biological target obtained through methods such as-

X-ray crystallography.

NMR spectroscopy.

Homologymodelling.



Selection of disease: determine the biochemical basis of disease process.

Target selection:

- Biochemical pathways could become abnormal and result in disease.
- Select a target at which to disrupt the biochemical process.
- Targets can be enzymes, receptors and nucleic acids.

Target validation:

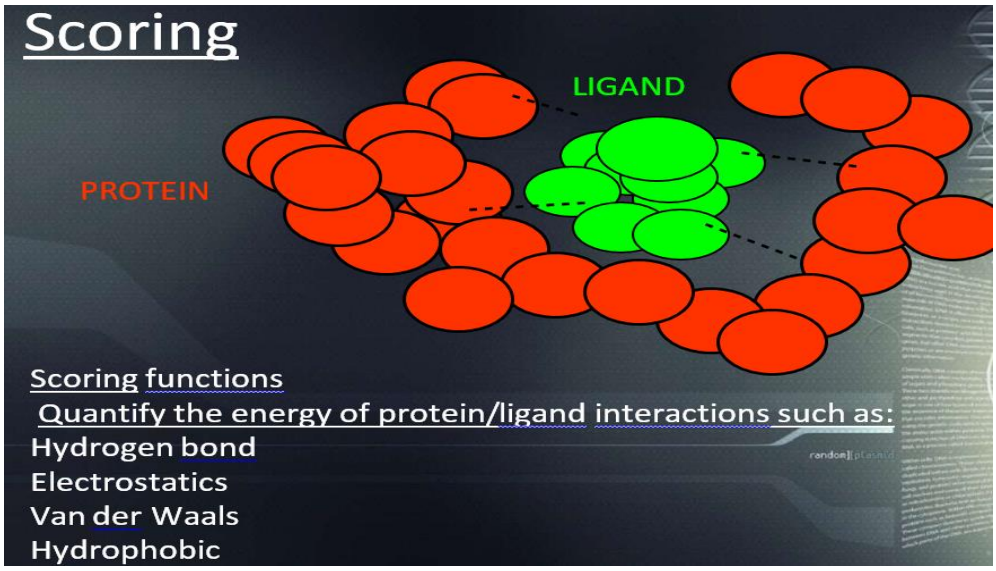
- Perform the protein BLAST for all genes/proteins with respect to Homo sapiens.
- Select the least matching molecule in human and again perform the BLAST.
- As the query sequence matched best, so we selected our target molecule and its structure can be obtained from RCSB AND PDB.

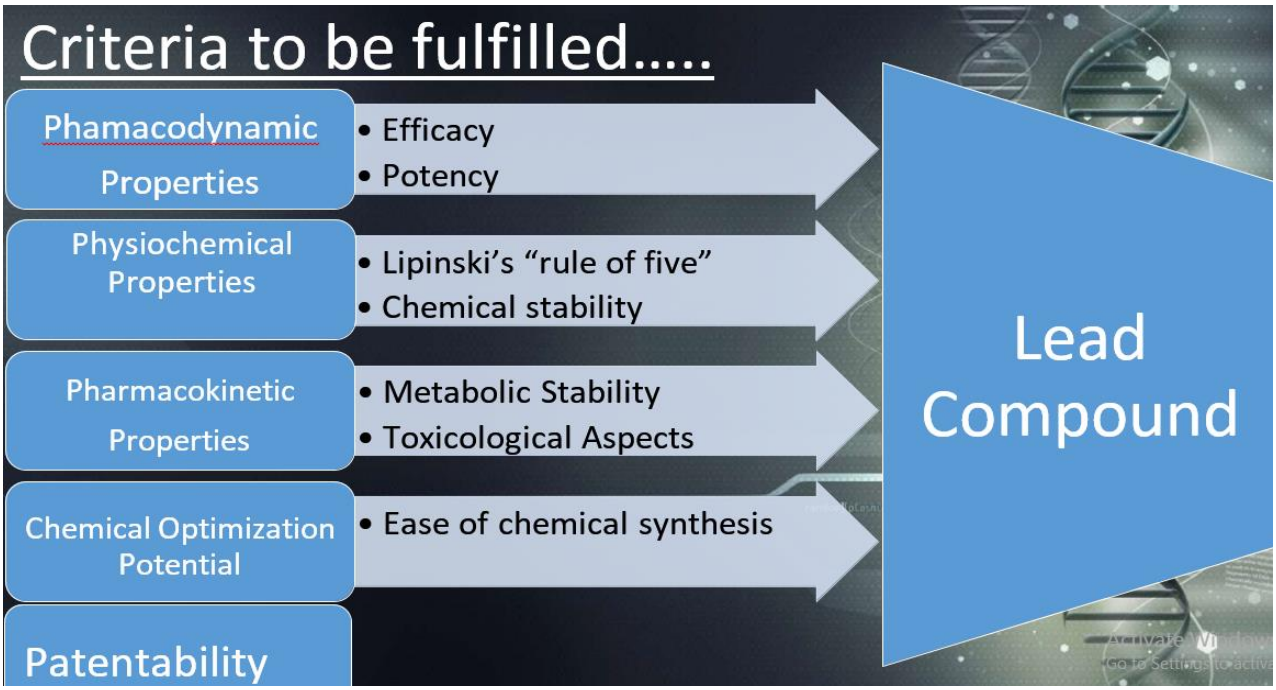
Structure determination:

Crystal structure of target protein can be taken from PDB database.

Selection of ligands or drugs:

- Also called as lead identification.
- High throughput screening of natural product and synthetic compound libraries is carried to screen out lead compounds.
- Docking- 3D structure of compound and target is docked.
- Scoring- scoring function evaluates complementarity.
- Selection- hits fulfill certain criteria and then selected as leads.


Scoring
LIGAND
PROTEIN
Scoring functions
Quantify the energy of protein/ligand interactions such as:
Hydrogen bond
Electrostatics
Van der Waals
Hydrophobic


Criteria to be fulfilled.....
Phamacodynamic Properties
• Efficacy
• Potency
Physiochemical Properties
• Lipinski's "rule of five"
• Chemical stability
Pharmacokinetic Properties
• Metabolic Stability
• Toxicological Aspects
Chemical Optimization Potential
• Ease of chemical synthesis
Patentability
Lead Compound

Lead optimization:

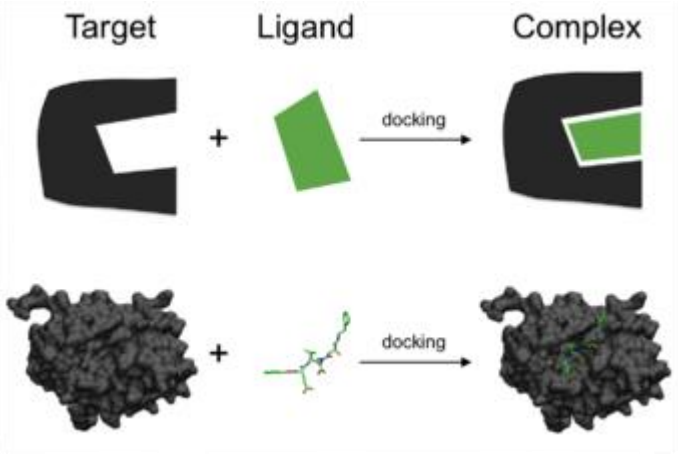
- Refining the 3-D structure of the lead.
- Technique is QSAR.

Preclinical and clinical development:

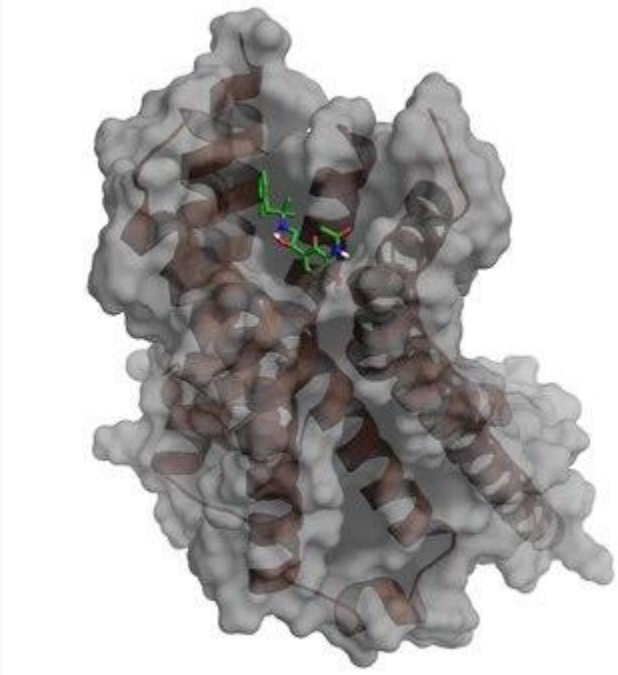
- **Preclinical trial**:
  . In vitro studies on animal.
  . Efficacy and pharmacokinetic information.
- **Clinical trial** :
  . 3 phases
  . Safety and efficacy on human beings.
- **File NDA:**
  . Document submitted to FDA for review.
  . FDA approval.

Molecular docking

In the field of molecular modeling, **docking** is a method which predicts the preferred orientation of one molecule to a second when bound to each other to form a stable complex. Knowledge of the preferred orientation in turn may be used to predict the strength of association or binding affinity between two molecules using, for example, scoring functions.



Schematic illustration of docking a small molecule ligand (green) to a protein target (black) producing a stable complex.



Docking of a small molecule (green) into the crystal structure of the beta-2 adrenergic G-protein coupled receptor.

The associations between biologically relevant molecules such as proteins, peptides, nucleic acids, carbohydrates, and lipids play a central role in signal transduction. Furthermore, the relative orientation of the two interacting partners may affect the type of signal produced (e.g., agonism vs antagonism). Therefore, docking is useful for predicting both the strength and type of signal produced.

Molecular docking is one of the most frequently used methods in structure-based drug design, due to its ability to predict the binding-conformation of small molecule ligands to the appropriate target binding site. Characterisation of the binding behaviour plays an important role in rational design of drugs as well as to elucidate fundamental biochemical processes.

QSAR (QUANTITY STRUCTURE ACTIVITY RELATIONSHIPS)

QSAR approach attempts to identify and quantify the physicochemical properties of a drug and to see whether any of these properties has an effect on the drug's biological activity by using a mathematical equation.

PHYSICOCHEMICAL PROPERTIES

• Hydrophobicity of the molecule

• Hydrophobicity of substituents
• Electronic properties of substituents
• Steric properties of substituents

WORKING OF QSAR

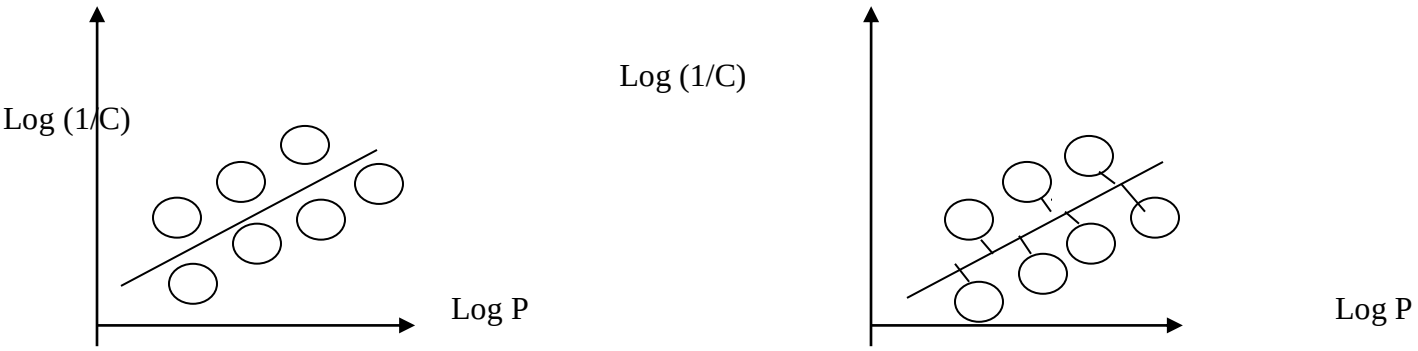
- A range of compounds is synthesized in order to vary one physicochemical property and to test it affects the bioactivity.

- A graph is then drawn to plot the biological activity on the y axis versus the physicochemical feature on the x axis.

- It is necessary to draw the best possible line through the data points on the graph. This done by procedure known as **linear regression analysis by the least square method.**

- If we draw a line through a set of data points will be scattered on either side of the line. The best line will be the one closest to the data points.

- To measure how close the data points are, vertical lines are drawn from each point.



HYDROPHOBICITY

Hydrophobic character of a drug is crucial to how easily it crosses the cell membrane and may also important in receptor interactions. Hydrophobicity of a drug is measured experimentally by testing the drugs relative distribution in octanol water mixture.

This relative distribution is known as partition coefficient.

$$\text{Partition Coefficient } P = \frac{[\text{Conc. Drug in Octanol }]}{[\text{Conc. of drug in water}]}$$

ELECTRONIC EFFECT

The electronic effect of various sustituents will clearly have an effect on drug ionisation and polarity. Have an effect on how easily a drug can pass through the cell membrane or how strongly it can interact with a binding site.

STEARIC FACTORS

The bulk, size and shape of a drug will influence how easily it can approach and interact with binding site. A bulky substituents may act like a shield and hinder the ideal interaction between a drug and its binding site. Bulky substituent may help to orient a drug properly for maximum binding and increase activity.

Pharmacokinetics Basics- Absorption, Distribution, Metabolism and Excretion

The four processes involved when a drug is taken are absorption, distribution, metabolism and elimination or excretion (ADME).

Pharmacokinetics is the way the body acts on the drug once it is administered. It is the measure of the rate (kinetics) of absorption, distribution, metabolism and excretion (ADME). All the four processes involve drug movement across the membranes. To be able to cross the membranes it is necessary that the drugs should be able dissolve directly into the lipid bilayer of the membrane; hence lipid soluble drugs cross directly whereas drugs that are polar do not.

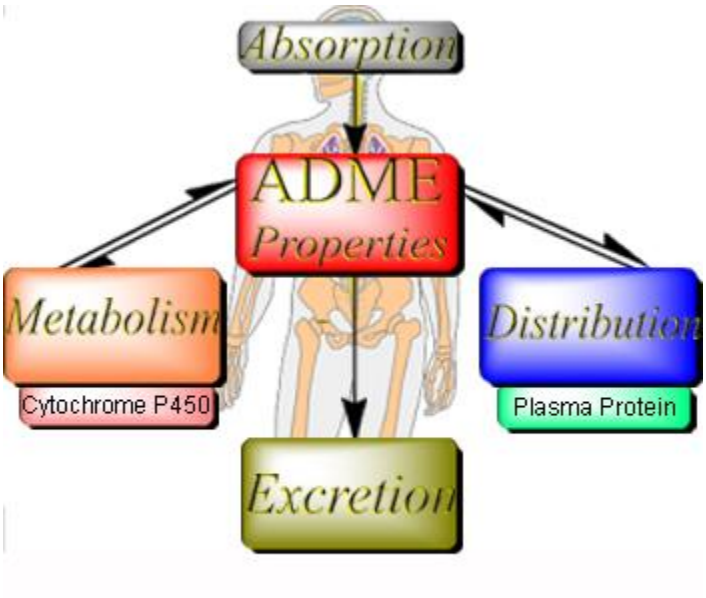


Figure showing the interplay between absorption, distribution, metabolism and excretion (ADME).

Absorption

Absorption is the movement of a drug from its site of administration into the blood. Most drugs are absorbed by passive absorption but some drugs need carrier mediated transport. Small molecules diffuse more rapidly than large molecules. Lipid soluble non – ionized drugs are absorbed faster. Absorption is affected by blood flow, pain stress etc.

Acidic drugs such as aspirin will be better absorbed in the stomach whereas basic drug like morphine will be absorbed better in the intestine. Most of the absorption of the drug takes place in the small intestine. Since the surface area of the stomach is much smaller than that of the intestine. Most of the drugs are absorbed in the small intestine since the amount of time that the drugs spend in the stomach is less and also the surface area of the stomach is small. If a basic drug is taken after a meal then the activity of the drug can be reduced whereas if an acidic drug is taken after a meal then the action of the can be noticed much more quickly, owing to the gastric absorption.

Distribution

Distribution is the movement of drugs throughout the body. Determined by the blood flow to the tissues, it is ability of the drug to enter the vasculature system and the ability of the drug to enter the cell if required.

- **Plasma Protein Binding**

  The blood stream has the ability to transport relatively insoluble substances. These substances are transferred by binding to the proteins which have a very amphipathic structure. The hydrophilic group renders the protein soluble in water and the lipophilic compounds are attracted to the lipophilic group and are loosely bound to the protein molecule hence protein bound. Most of the drugs travel in the

plasma are partly in solution and partly bound to the plasma protein. The bound drug is inactive and the unbound drug is active. The ratio of bound to the unbound drug varies. Binding is reversible. Generally acidic drugs bind to albumin and basic drugs to α1 – acid glycoprotein.

- **Tissue Distribution**

  After absorption most drugs are distributed in the blood to the body tissue where they have their effect. The degree to which the drug is likely to accumulate in the tissue is dependent on the lipophilicity and local blood flow to the tissue. Highly perfused organs receive most of the drugs.

- **The role of the liver in drug distribution**

  After the drug is absorbed by the GI tract, it is taken up by the part of the bloodstream called the hepatic portal system. Most of the drugs are absorbed into this system except for the lipids which are absorbed into the lymphatic system and then delivered into the blood by the thoracic duct into the superior vena cava.

Metabolism or Biotransformation

It is the process of transformation of a drug within the body to make it more hydrophilic so that it can be excreted out from the body by the kidneys. This needs to be done since drugs and chemicals are foreign substances in our body. If the drug continues to be in the lipohilic state and is going to be filtered by the glomerulus then it will be reabsorbed and remain in the body for prolonged periods. Hence metabolism deals with making the drug more hydrophilic such that it can be excreted out from the body. In some cases the metabolites can be more active than the drug itself e.g. anxiolytic benzodiazepines.

Excretion

Excretion is the removal of the substance from the body. Some drugs are either excreted out unchanged or some are excreted out as metabolites in urine or bile. Drugs may also leave the body by natural routes such as tears, sweat, breath and saliva. Patients with kidney or liver problem can have elevated levels of drug in the system and it may be necessary to monitor the dose of the drug appropriately since a high dose in the blood can lead to drug toxicity.

Pharmacodynamics

Pharmacodynamics describes the action of drug on body and influence of drug concentration on magnitude of response.

Most of drugs exerts effect by interacting with receptors.

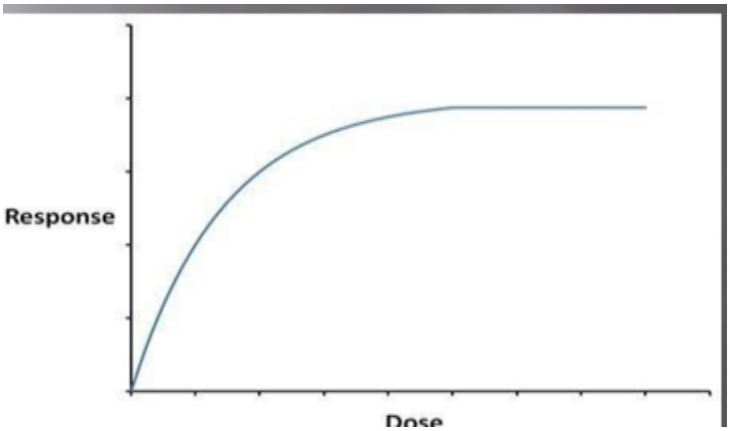
An agent that can bind to receptor and produces biological response is called as agonist.

Drug response curve-

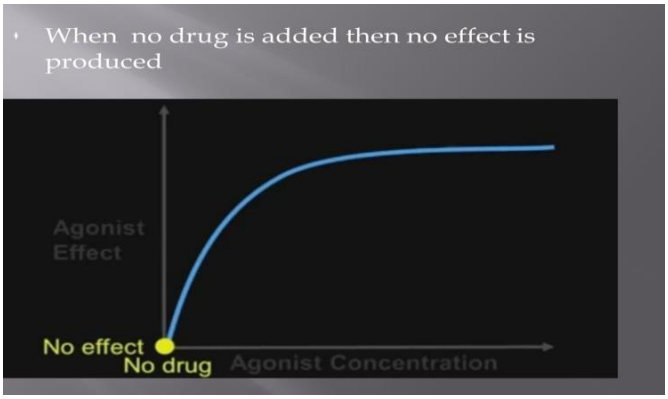
The interaction between drug and its receptor can be described by a curve called as drug response curve.

On vertical axis there is response of drug.

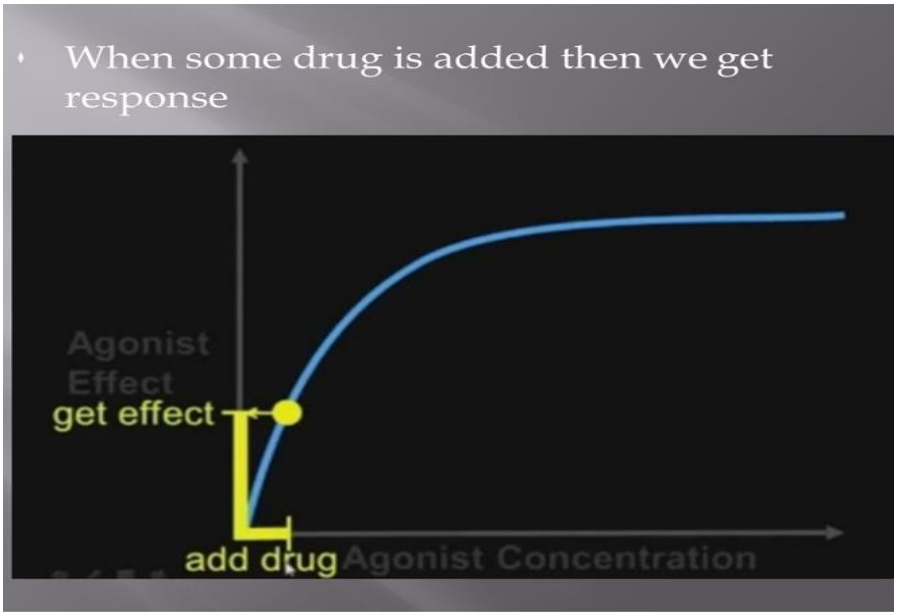
On horizontal axis there is concentration of dose.



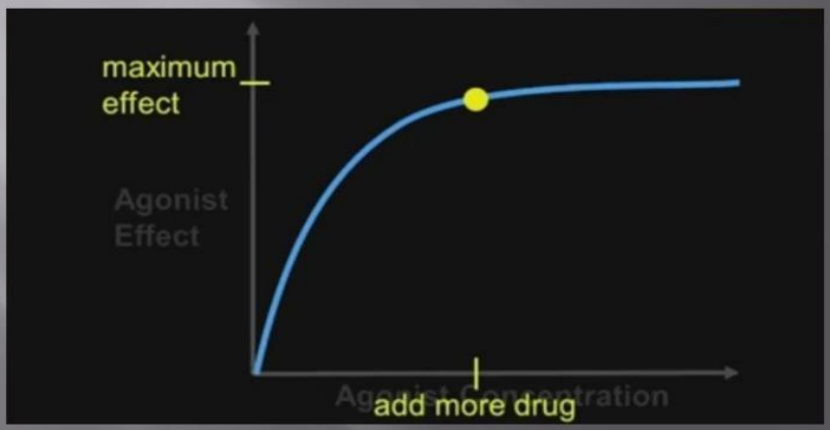
The magnitude of drug effect depends on drug concentration at receptor site and this availability and concentration of drug at receptor is determined by both dose of drug administered and by drug pharmacodynamics (ADME).



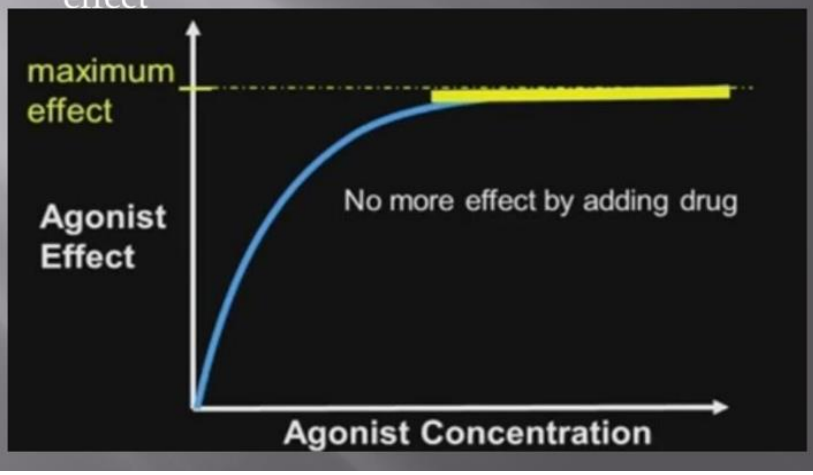
- When some drug is added then we get response



- By adding more drug we get maximum response

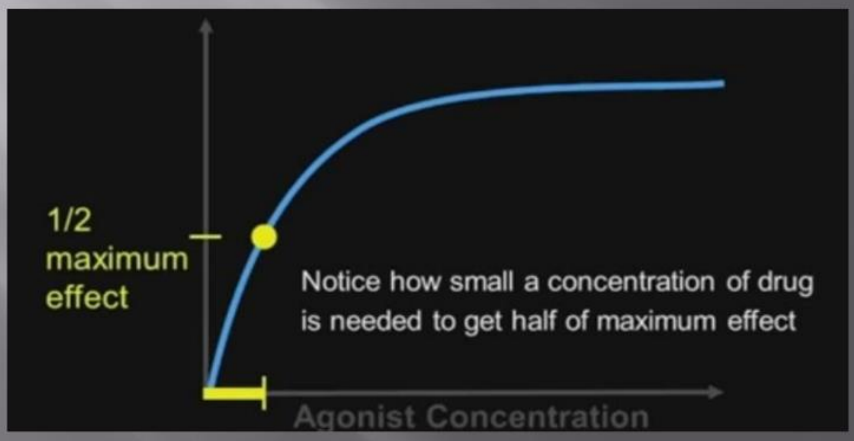


- A time will reach on which adding more drug will not cause increase in response because we have reached to maximum effect



- **EC50**

Is concentration of drug needed to get half of the maximum effect



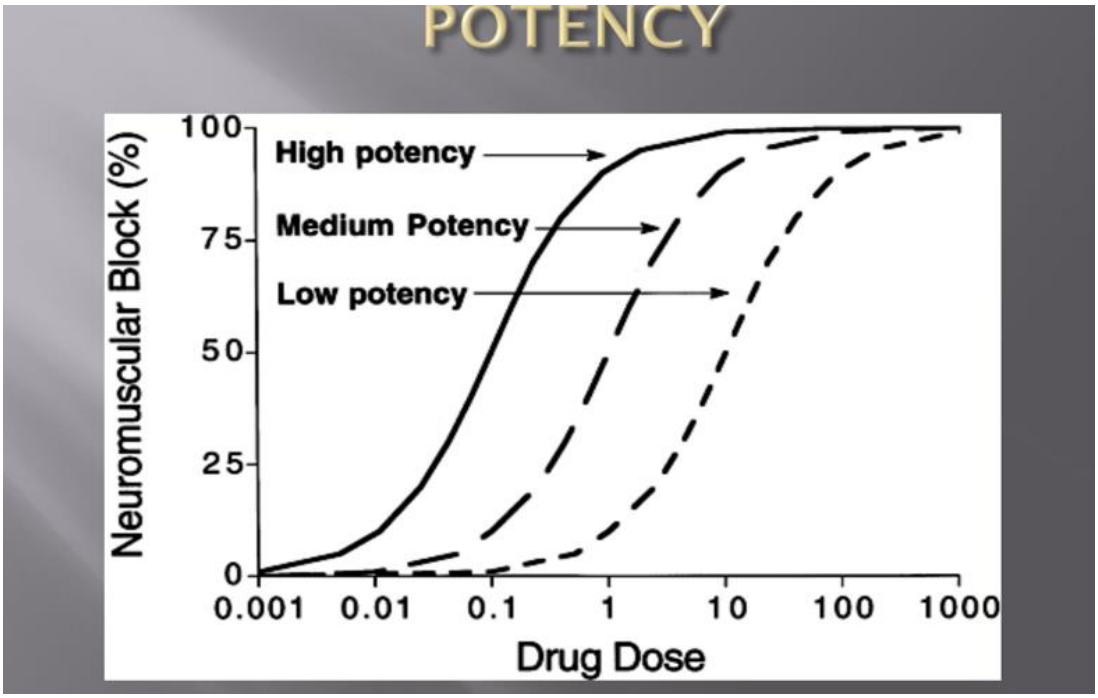
Two important properties of drug are –

Efficacy

Potency

Potency-

Amount of drug necessary to produce maximum effect is the potency of drug.



Efficacy-

Ability of a drug to elicit a response when it interacts with a receptor.

Dependent on – no of drug-receptor complexes formed

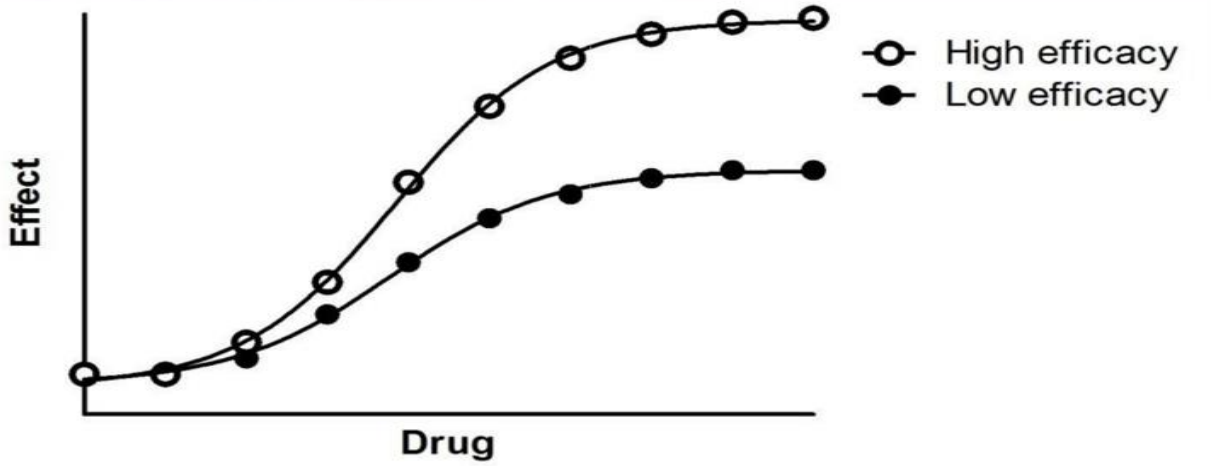
- Efficiency of the coupling of receptor activation to cellular responses.

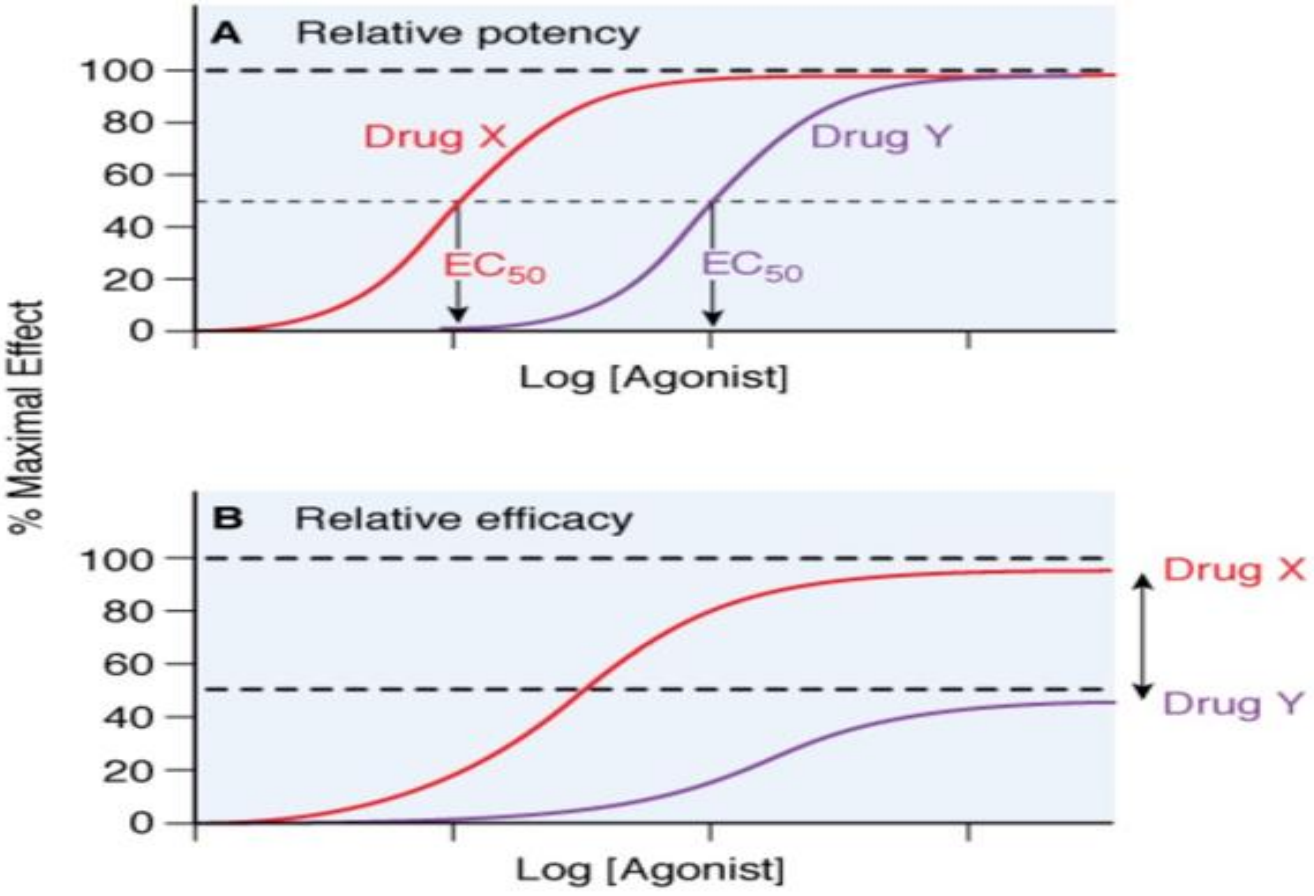
Maximal efficacy of a drug assumes that all receptors are occupied by the drug and if more drugs are added, no additive response will be observed.

Maximal response (efficacy) is more important than drug potency.

A drug with greater efficacy is more therapeutically beneficial than the one that is more potent.

EFFICACY


High efficacy
Low efficacy
Effect
Drug

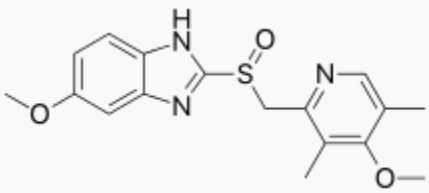

A Relative potency
100
80
60
40
20
0
Drug X
Drug Y
EC50
EC50
Log [Agonist]
% Maximal Effect
B Relative efficacy
100
80
60
40
20
0
Drug X
Drug Y
Log [Agonist]

Lipinski's rule of five-

also known as **Pfizer's rule of five** or simply the **rule of five** (**RO5**), is a rule of thumb to evaluate druglikeness or determine if a chemical compound with a certain pharmacological or biological activity has chemical properties and physical properties that would make it a likely orally active drug in humans. The rule was formulated by Christopher A. Lipinski in 1997, based on the observation that most orally administered drugs are relatively small and moderately lipophilic molecules.[1][2]

The rule describes molecular properties important for a drug's pharmacokinetics in the human body, including their absorption, distribution, metabolism, and excretion ("ADME"). However, the rule does not predict if a compound is pharmacologically active.

The rule is important to keep in mind during drug discovery when a pharmacologically active lead structure is optimized step-wise to increase the activity and selectivity of the compound as well as to ensure drug-like physicochemical properties are maintained as described by Lipinski's rule. Candidate drugs that conform to the RO5 tend to have lower attrition rates during clinical trials and hence have an increased chance of reaching the market.



Omeprazole is a popular drug that conforms to Lipinski's rule of five.

Components of the rule

Lipinski's rule states that, in general, an orally active drug has no more than one violation of the following criteria:

- No more than 5 hydrogen bond donors (the total number of nitrogen–hydrogen and oxygen–hydrogen bonds)
- No more than 10 hydrogen bond acceptors (all nitrogen or oxygen atoms)
- A molecular mass less than 500 daltons
- An octanol-water partition coefficient (log P) that does not exceed 5

Note that all numbers are multiples of five, which is the origin of the rule's name. As with many other rules of thumb, such as Baldwin's rules for ring closure, there are many exceptions.

Variants

In an attempt to improve the predictions of drug-likeness, the rules have spawned many extensions, for example the Ghose filter:

- Partition coefficient log P in −0.4 to +5.6 range

- Molar refractivity from 40 to 130
- Molecular weight from 180 to 480
- Number of atoms from 20 to 70 (includes H-bond donors [e.g. OHs and NHs] and H-bond acceptors [e.g. Ns and Os])

Veber's Rule further questions a 500 molecular weight cutoff. The polar surface area and the number of rotatable bonds has been found to better discriminate between compounds that are orally active and those that are not for a large data set of compounds in the rat.In particular, compounds which meet only the two criteria of:

- 10 or fewer rotatable bonds and
- Polar surface area no greater than 140 $Å^2$

are predicted to have good oral bioavailability.

Lead-like

During drug discovery, lipophilicity and molecular weight are often increased in order to improve the affinity and selectivity of the drug candidate. Hence it is often difficult to maintain drug-likeness (i.e., RO5 compliance) during hit and lead optimization. Hence it has been proposed that members of screening libraries from which hits are discovered should be biased toward lower molecular weight and lipophility so that medicinal chemists will have an easier time in delivering optimized drug development candidates that are also drug-like. Hence the rule of five has been extended to the **rule of three** (RO3) for defining **lead-like** compounds.

A rule of three compliant compound is defined as one that has:

- octanol-water partition coefficient log P not greater than 3
- molecular mass less than 300 daltons
- not more than 3 hydrogen bond donors
- not more than 3 hydrogen bond acceptors
- not more than 3 rotatable bonds

Pharmacogenomics :-Pharmacogenomics is the study of the role of the genome in drug response. Its name (pharmaco + genomics) reflects its combiningof pharmacology and genomics. Pharmacogenomics analyzes how the genetic makeup of an individual affects his/her response to drugs. It deals with the influence of acquired and inherited genetic variation on drug response in patients by correlating geneexpression or single-nucleotide polymorphisms with pharmacokinetics (drug absorption, distribution, metabolism, and elimination) and pharmacodynamics (effects mediated through a drug's biological targets). The term pharmacogenomics is often used interchangeably with pharmacogenetics. Although both terms relate to drug response based on genetic influences, pharmacogenetics focuses on single drug-gene interactions, while pharmacogenomics encompasses a more genome-wide association

approach, incorporating genomics and epigenetics while dealing with the effects of multiple genes on drug response.Pharmacogenomics aims to develop rational means to optimize drug therapy, with respect to the patients' genotype, to ensure maximum efficiency with minimal adverse effects. Through the utilization of pharmacogenomics, it is hoped that pharmaceutical drug treatments can deviate from what is dubbed as the "one-dose-fits-all" approach. Pharmacogenomics also attempts to eliminate the trial-and-error method of prescribing, allowing physicians to take into consideration their patient's genes, the functionality of these genes, and how this may affect the efficacy of the patient's current or future treatments (and where applicable, provide an explanation for the failure of past treatments). Such approaches promise the advent of precision medicine and even personalized medicine, in which drugs and drug combinations are optimized for narrow subsets of patients or even for each individual's unique genetic makeup. Whether used to explain a patient's response or lack thereof to a treatment, or act as a predictive tool, it hopes to achieve better treatment outcomes, greater efficacy, minimization of the occurrence of drug toxicities and adverse drug reactions (ADRs). For patients who have lack of therapeutic response to a treatment, alternative therapies can be prescribed that would best suit their requirements. In order to provide Pharmacogenomic recommendations for a given drug, two possible types of input can be used: genotyping or exome or whole genome sequencing.Sequencing provides many more data points, including detection of mutations that prematurely terminate the synthesized protein (early stop codon).

(Following Paper ID and Roll No. to be filled in your Answer Books)

Paper ID : 154613 Roll No. | | | | | | | | | | |

B. TECH.

Theory Examination (Semester-VI) 2015-16

BIOINFORMATICS-II

Time : 3 Hours *Max. Marks : 100*

Section-A

1. Attempt all parts. All parts carry equal marks. Write answer of each part in short. (2 × 10 = 20)

(a) What is genomic sequence annotation? Give some tools for Gene and ORF prediction.

(b) What is Microarray? How bioinformatics is applied to analyse microarray data?

(c) What is machine learning? Name some machine learning approaches.

(d) What is a Decision Tree? Give example.

(e) What is computer simulation?

(f) Describe the relation between statistics and machine learning.

(g) Explain the technique of Document clustering.

(h) What is Lipinski's rule of five in insilico drug designing?

(i) What is the diffemce between Parametric and Non-Parametric tests?

(j) Explain Perl? What are Arrays, Hashes in Perl ?

Section-B

2. **Attempt any five parts. All parts cany equal marks :** (10×5=50)

(a) What is simulated annealing ?

(b) How is Artificial Neural Network helpful in solving biological problems?

(c) Discuss and describe the Genetic Algorithm.

(d) Explain Natural Language Processing. Discuss its major areas.

(e) Describe computer simulation techniques and its types.

(f) Explain Pharmacodynamics (Efficacy & Potency) & Pharmacokinetics (ADME).

(g) Explain the methods of clustering (Hierarchical and K-mean).

(h) Describe Hidden Markov Model and one of its application.

Section-C

Note : Attempt any two questions from this section. (15×2=30)

3. Describe and discuss some molecular biology techniques and their inference problems solved by the help of bioinformatics.

4. What is force field in molecular modeling? How is it helpful in study of molecular dynamic simulation?

5. What is in silico drug designing? Explain Ligand and Structure based drug designing.