Introduction to bioinformatics

(Gene prediction, ORF Finder, Primer 3, and Phylogenetic prediction)

Course Code –BOTY 4204

Course Title- Techniques in plant sciences , biostatistics and bioinformatics

By – Dr. Alok Kumar Shrivastava
Department of Botany
Mahatma Gandhi Central University, Motihari

Gene Prediction

- ✓ Gene prediction by computational methods for finding the location of protein coding regions is one of the essential issues in bioinformatics.
- ✓ The gene sequencing of a gene is productive only when it is analysed and predicted correctly.
- ✓ Gene prediction is carried out to identify the structure of genes in order to differentiate protein coding genes from non coding regions, and to identify promoters and other regulatory elements.
- ✓ Gene prediction basically means locating genes along genome. Also called gene finding, it refers to the process of identifying the regions of genomic DNA that encode genes.
- ✓ This includes protein coding genes, RNA genes and other functional elements such as the regulatory genes.

Importance of gene prediction

✓Helps to annotate large, contiguous sequences.

✓It provides information on the evolution of genes, speciation and evolution of species.

✓It gives an understanding of the structure and function of genomes of different organisms.

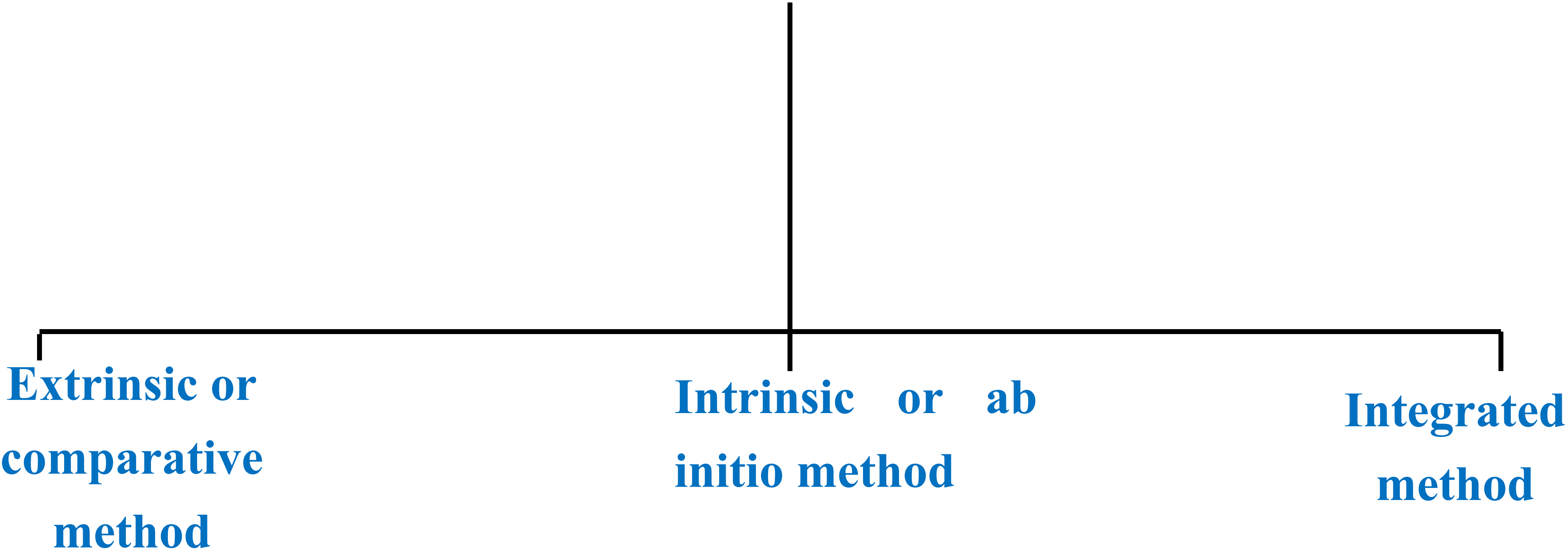
✓Distinguish between coding and non coding regions of a genome.

✓Aids in the identification of fundamental and essential elements of genome such as functional genes, intron, exon, splicing sites, gene encoding known proteins, motifs, EST, etc.

✓Describe individual genes interms of their function.

✓It has vast application in structural genomics, metabolomics, proteomics, transcriptomics, functional geenomics, genome studies and other genetic related studies including genetic disorder detection, treatment and prevention.

Methods of gene prediction



Extrinsic or comparative method

Intrinsic or ab initio method

Integrated method

Extrinsic or comparative method

- ✓ It is a method based on sequence similarity searches.
- ✓ It is a conceptual simple approach that is based on finding similarity ingene sequences between ESTs (expressed sequence tags), proteins, or other genomes to the input genome.
- ✓ This approach is based on the assumption that functional regions (exons) are more conserved evolutionary than non functional regions (intergenic or intronic regions)
- ✓ Once there is similarity between a certain genomic region and an EST, DNA, or protein, the similarity information can be used to infer gene structure or function of that region.

Continue……………

- Local alignment and global alignment are two methods based on similarity searches. The most common local alignment tool is the BLAST family of programs, which detects sequence similarity to known genes, proteins, or ESTs.

- Two more types of software, PROCRUSTES and GeneWise, use global alignment of a homologous protein to translated ORFs in a genomic sequence for gene prediction.

- Anew heuristic method based on pairwise genome comparision has been implemented in the software called CST finder.

Intrisic or Ab- initio prediction method

- This method uses known properties of coding and non coding sequences like open reading frame and coding statistics (ORF length, codon usage, GC content, etc.) for gene prediction.
- It uses gene structure as a template to detect genes.
- Ab initio gene prediction rely on two types of sequence information ; signal sensor and content sensor.
- Signal sensors refer to short sequences motifs, such as splice sites, branch points, polypyrimidine tracts, start codons and stop codons.
- Content sensors refer to the patterns of codon usage that are unique to a species, and allow coding sequence to be distinguished from the surrounding non-coding sequences. Exon detection must rely on the content sensors.

continue…………

- Many algorithms are applied for modeling gene structure, such as, Dynamic Programming, Linear discriminant analysis, Linguist methods, Hidden Markov Model and Neural Network.
- Based on these models, a great number of ab initio gene prediction programs have been developed. Some of the frequently used ones are Gene ID, FGENESH, GeneParser, GlimmerM, GENSCAN etc.

Integrated approach

- It is combines both intrinsic and extrinsic methods. It combines the calculations of several programs for better prediction.

Tools for gene prediction

- ORF Finder
- GeneBee
- GeneMark
- GENESCAN
- Promoter 2.0 prediction server
- Gene Finder

ORF Finder

- In molecular genetics, an open reading frame is the part of a reading frame that has the ability to be translated. An ORF is a continuous stretch of codons that begins with a start codon and ends at a stop codon. An ATG codon within the ORF may indicate where translation starts. In other words we can say that the region of a nucleotide that starts from an initiation codon and ends with a stop codon is called ORF.
- The CDS(coding sequence) is the actual region of DNA that is translated to form proteins while the ORF may contain introns as well. The CDS refers to those nucleotides (concatenated exons) that can be divided into codons which are actually translated into amino acids by the process of translation.
- ORF finder is a program or graphical analysis tool available at NCBI website which searches for open reading frames(ORFs) in the DNA sequence you enter. the program or tool returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments. This tool identifies all open reading frames using the standard or alternative genetic codes.

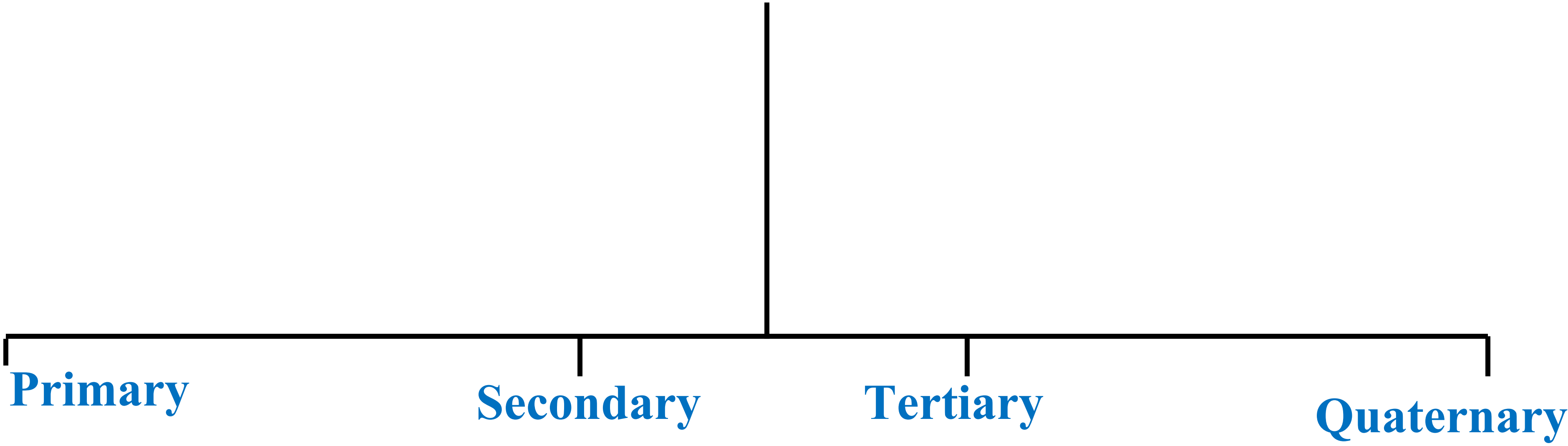
Primer 3

- Primer3 is a widely used program for designing PCR primers. PCR is an essential and ubiquitous tool in genetics and molecular biology. Primer3 can also design hybridization probes and sequencing primers.
- Primer is a short (18-24nt) single stranded oligonucleotides used for DNA synthesis. It is complementary to DNA sequences that flank the target region of interest. The purpose of primer is to provide a free 3’-OH group to which the DNA polymerase can add dNTPs.

Protein structure and prediction

- Proteins are made up of amino acids molecules. All amino acids contain the elements carbon, hydrogen, oxygen, and nitrogen. Some amino acids also have sulphur and phosphorus, and trace elements such as iron or copper. There are 20 different amino acids and they can be linked together in different combinations or sequences to form different proteins. Amino acids are covalently linked through peptide bonds to form linear polymers called proteins or peptides. On synthesis, proteins spontaneously fold into three dimensional structures, the shape of which determines their biological function.

Structure of proteins



Primary

Secondary

Tertiary

Quaternary

Primary structure

✓The primary structure is the sequence of the amino acids or exact specification of its atomic composition and the chemical bonds connecting those atoms. It start from amino-terminal(N) end and terminate at the carboxyl- terminal (C)end.

✓Primary structure prediction tools allow the computation of various physical and chemical parameters for a given protein sequence. It compute the molecular weight, amino acid composition, atomic composition, extinction coefficient, estimated half life, instability index, aliphatic index and grand average hydropathicity.

✓ProtParam, ProtScale, RandSeq, etc. are tools for primary structure prediction of proteins.

Secondary structure

- The secondary structure is the way in which linear chain of amino acids orients itself stabilised by hydrogen bonds. The most common secondary structures are alpha – helices and beta- sheets. Different sequences of amino acids form different secondary structure elements.
- For example if a sequence comprising methionine, alanine, leucine, glutamate and lysine may prefer to adopt alpha-helical conformations in proteins; proline and glycine are commonly found in turns; isoleucine, valine and threonine prefer to adopt beta-sheet conformations.
- Secondary structure derived from these sequences confer unique properties and functions.
- SOPMA, PHD, etc. are the common tools for secondary structure prediction of protein.

Tertiary structure

The tertiary structure is the three dimensional structure of the protein which it assumes by a folding process. It is stabilised by disulphide bonds and non covalent interactions like hydrogen bonding, hydrophobic forces, van der waals forces, ionic interactions, etc.

The different strategies involved in 3D protein structure prediction; comparative modelling or homology modelling, Ab initio method, and Swiss PDB viewer and MODELLER

The predicted structure can be validated using PROCHECK, WHATCHECK, etc.

The Ramachandran plot for the 3D structure is used to finally confirm stability based on the free energy of the protein structure.

Quaternary structure

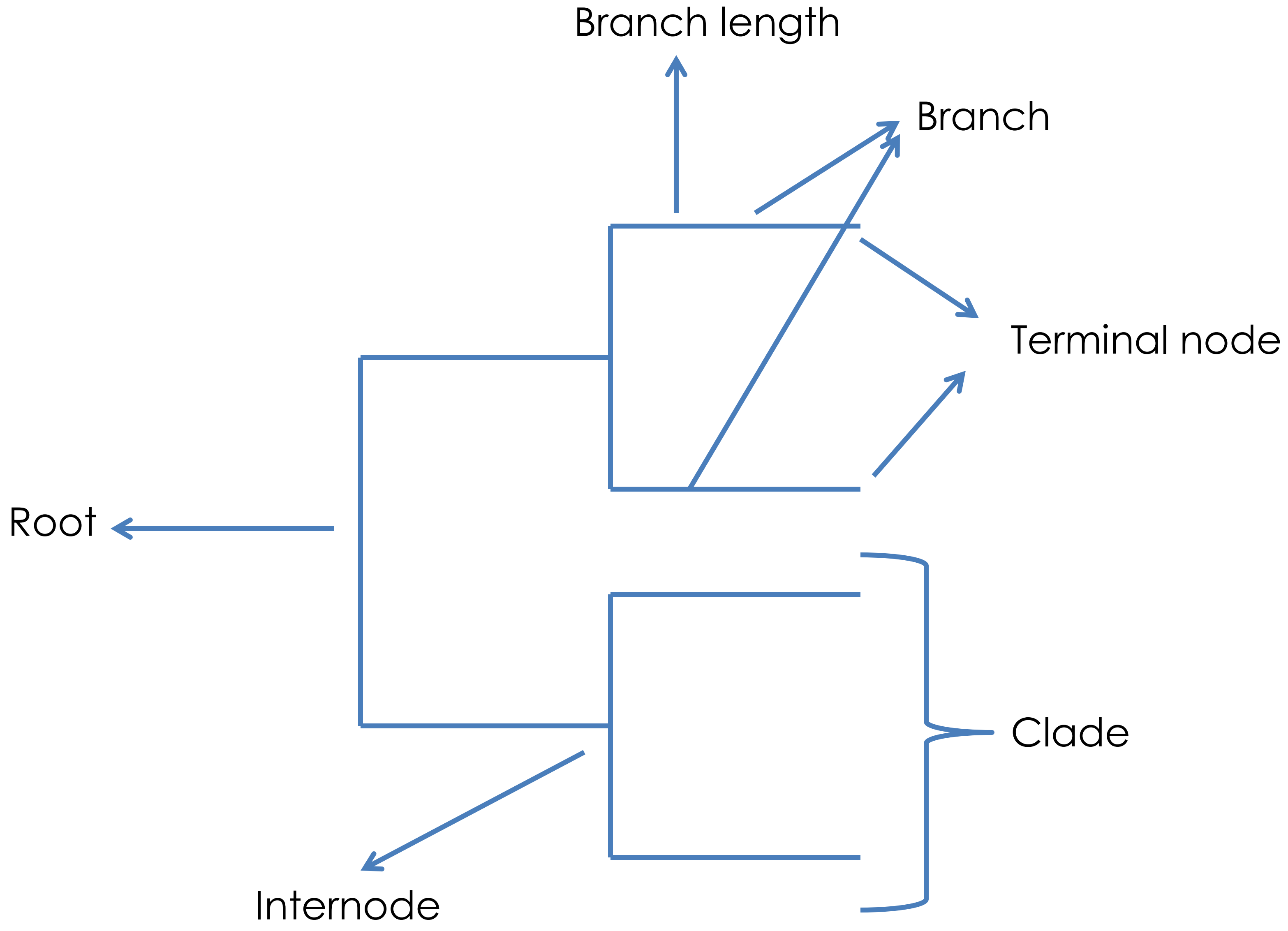
- Several individual peptide or protein chains cluster into a final definite shape- the quaternary protein structure. Hydrogen bonding, salt bridges, and disulphide bonds hold the various chains in a particular geometry. There are two main categories of proteins with quaternary structure; fibrous (silk proteins and keratins) and globular (insulin, haemoglobin and most enzymes).

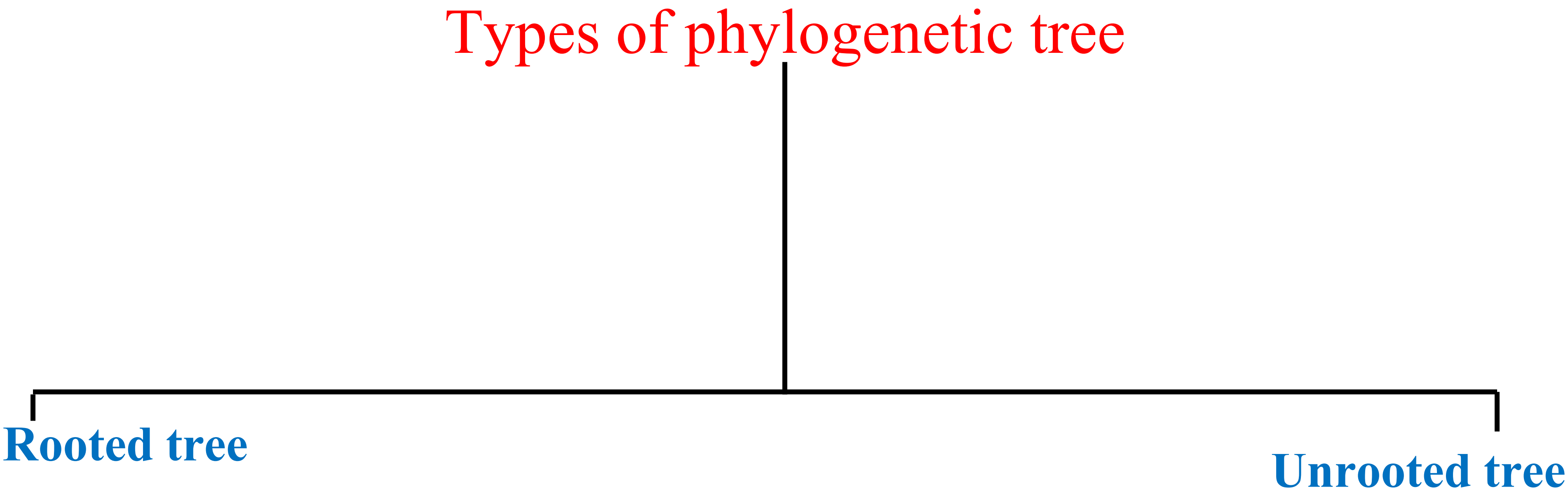
Phylogeny

- ✓ Phylogeny is the study of patterns of similarity among living organisms.
- ✓ Phylogenetics can be defined as the field of biology dealing with the characterisation of relationships among various live components on earth. This study starts with collecting data and analysing them, and extends to the interpretation of those results to give a new meaning.
- ✓ Phylogenetic systematics deals with identifying and understanding the evolutionary relationship among extant(living) and extinct(dead) organisms on earth.
- ✓ Cladistics is the study of the pathway of evolution.
- ✓ Phenetics is the study of the relationships among a group of organisms on the basis of the degree of similarity between them(molecular, phenotypic and anatomical).

Phylogenetic trees

- A phylogenetic tree is a graphical representation of the evolutionary relationship of genes/taxa/sequences is called a phylogenetic tree. A phylogenetic tree is constructed by nodes representing the taxonomic unit, branches representing the relationship between the taxonomic units and roots representing the ancestor sequence. Cladistic is the branching pattern of the tree.


Branch length
Branch
Terminal node
Root
Clade
Internode



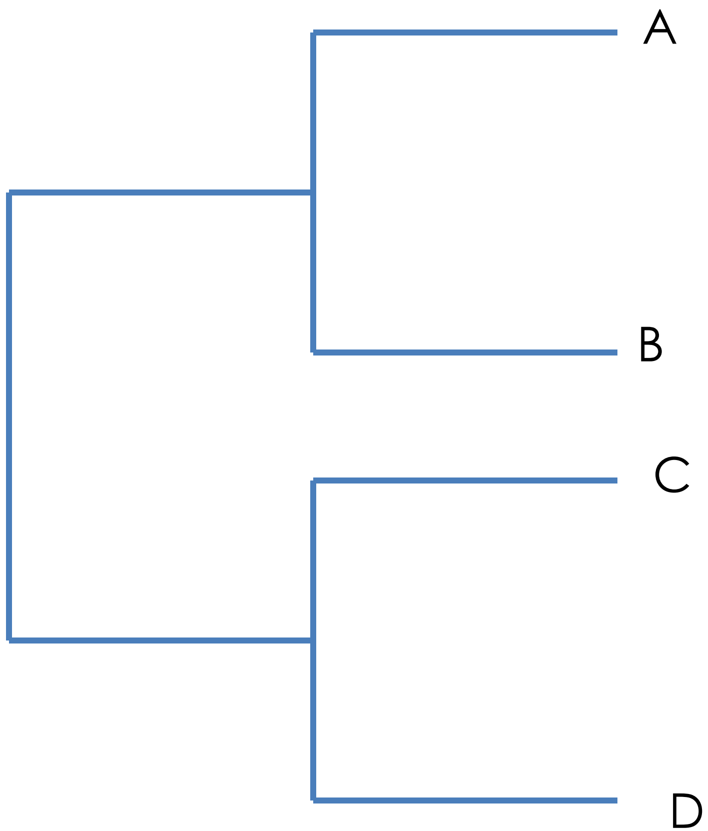
Types of phylogenetic tree

Rooted tree

Unrooted tree

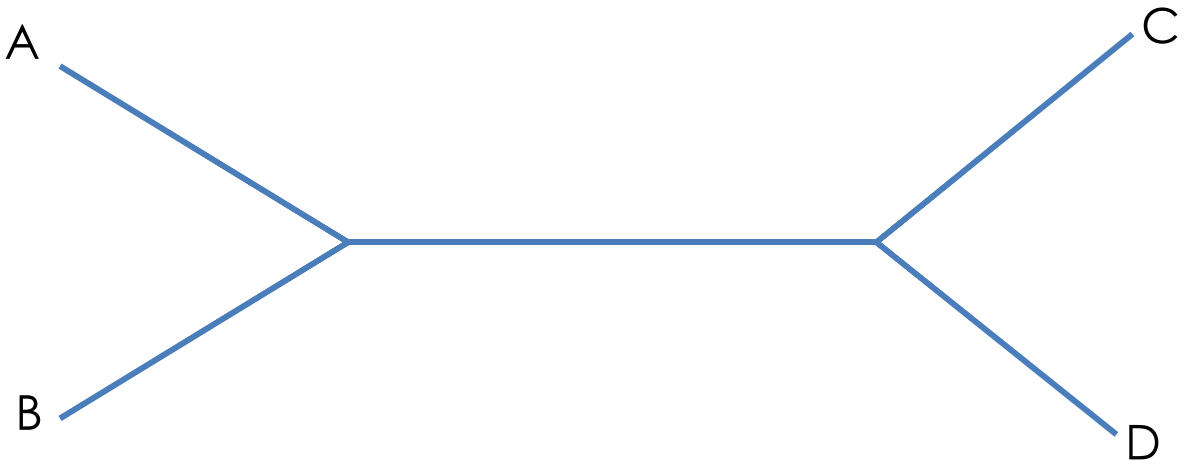
Rooted tree

Rooted tree is a tree where a node is defined as the root. A rooted tree is based on the molecular clock and provides both the evolutionary path as well as the relationship among the species. the resultant tree is called as cladogram.



Unrooted tree

- It specifies the relationship among species without identifying a common ancestor or evolutionary path. The resultant tree is called a phenogram.



Terminology related to tree

- Node-it is the taxonomic unit (existing species or an ancestor) also called OTU (operational taxonomic units)
- Distance scale- represent the number of differences between the sequences.
- Clade- Clade can be explained as a group of two or more taxonomic units coming from the same common ancestor.
- Branch- describes the relationship between the trees in terms of descent and ancestry.
- Degree- it is number of adjacent branches present in an internal node.

Continue…………………..

Branch length- It suggests the number of changes that have occurred in the branch.

Phenogram- It is classified as tree like representation expressing phenetic relationships.

Phylogram- It is a phylogenetic tree that explicitly represent the number of mutations or character changes through its branch lengths.

Chronogram- It is a phylogenetic tree that explicitly represents evolutionary time through its branch lengths.

Polytomy- It is classified as a node with a degree greater than three(one ancestor having more than two immediate descendants). a fully resolved tree does not have polytomy.

Continue………………..

Cladogram

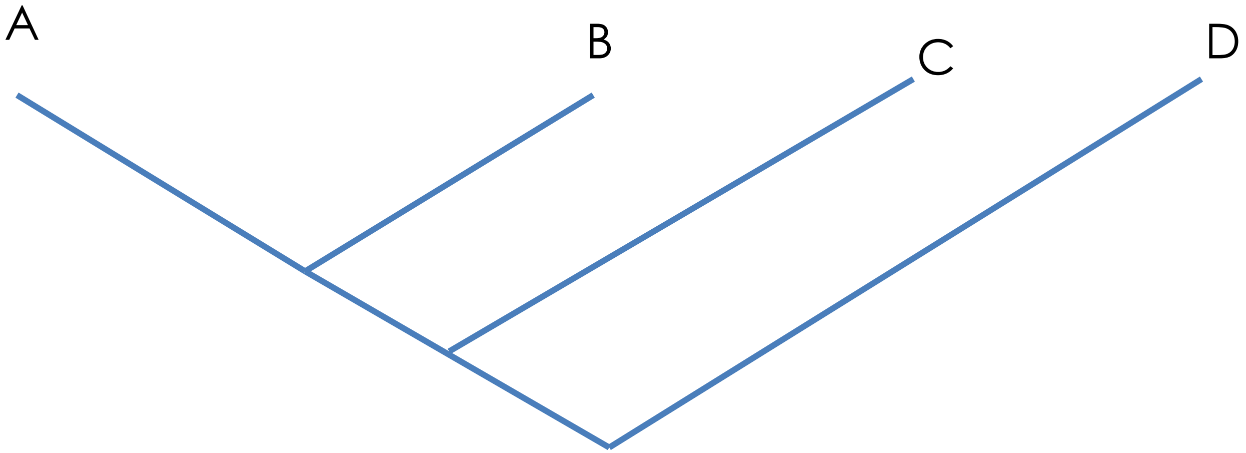
✓It is defined as the topology of a rooted phylogenetic tree. Evolutionary trees represent clade, a group of organisms that includes an ancestor and all descendents of that ancestor. The simplest statement about an evolutionary relationship cladogram is a tree formed using cladistic methods.

✓This type of tree only represent branching pattern , that is , its branch lengths do not represent time.

✓ at the y junction, that is, at the new branch , novel characters of evolutionary origin are used to split one group from the rest .

✓Cladograms emphasise the sequence or order in which derived characters arise from a central phylogenetic tree.

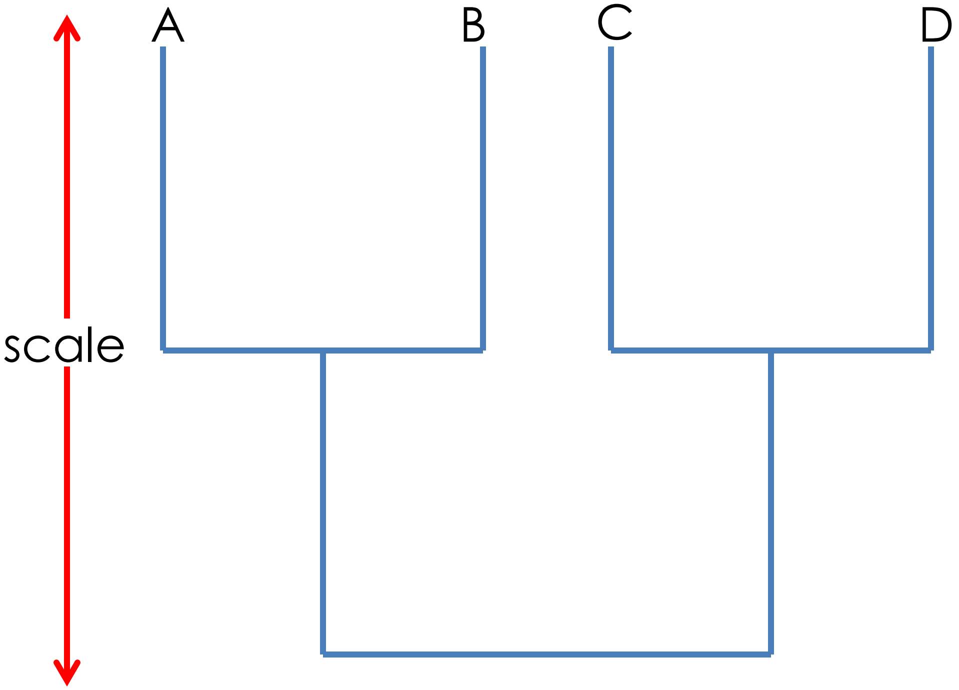
✓It does not indicate how strong or profound the derived character is, and its evolutionary importance.



Cladogram

Dendrogram or ultrametric tree

- Dendrogram is a broad term for the diagrammatic representation of a phylogenetic tree. It is an additive tree , a generalisation of ultrameric trees, where the tips of the tree are equidistant from the root of the tree. we can explain it as the number of mutations in the additive tree and is assumed to be proportional to the chronological distance of a node to the ancestor; it also assumed that the mutations take place with the same rate in all paths. An additive tree contains information about branch lengths. The direction along the y axis in additive trees represents the amount of change or time.

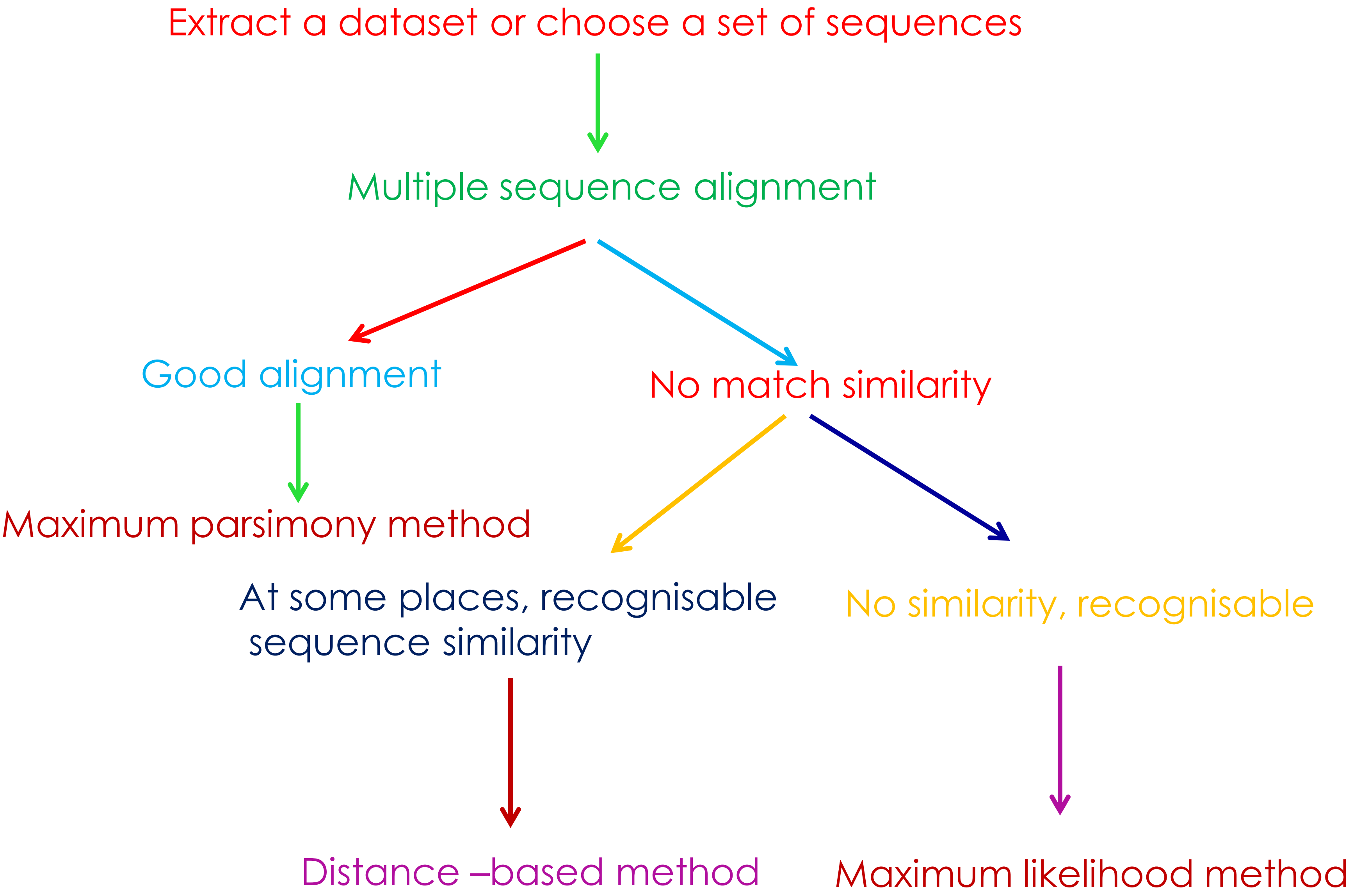


Dendrogram

Steps to construct phylogenetic tree

- Extract a database (sequences) and build a data model (alignment method).
- Consider sequence variation and determine the substitution model.
- Build the tree using distance-based or character-based methods.
- Construct the resultant tree by character evaluation.

Steps for selection of method to build a tree



Extract a dataset or choose a set of sequences

Multiple sequence alignment

Good alignment

No match similarity

Maximum parsimony method

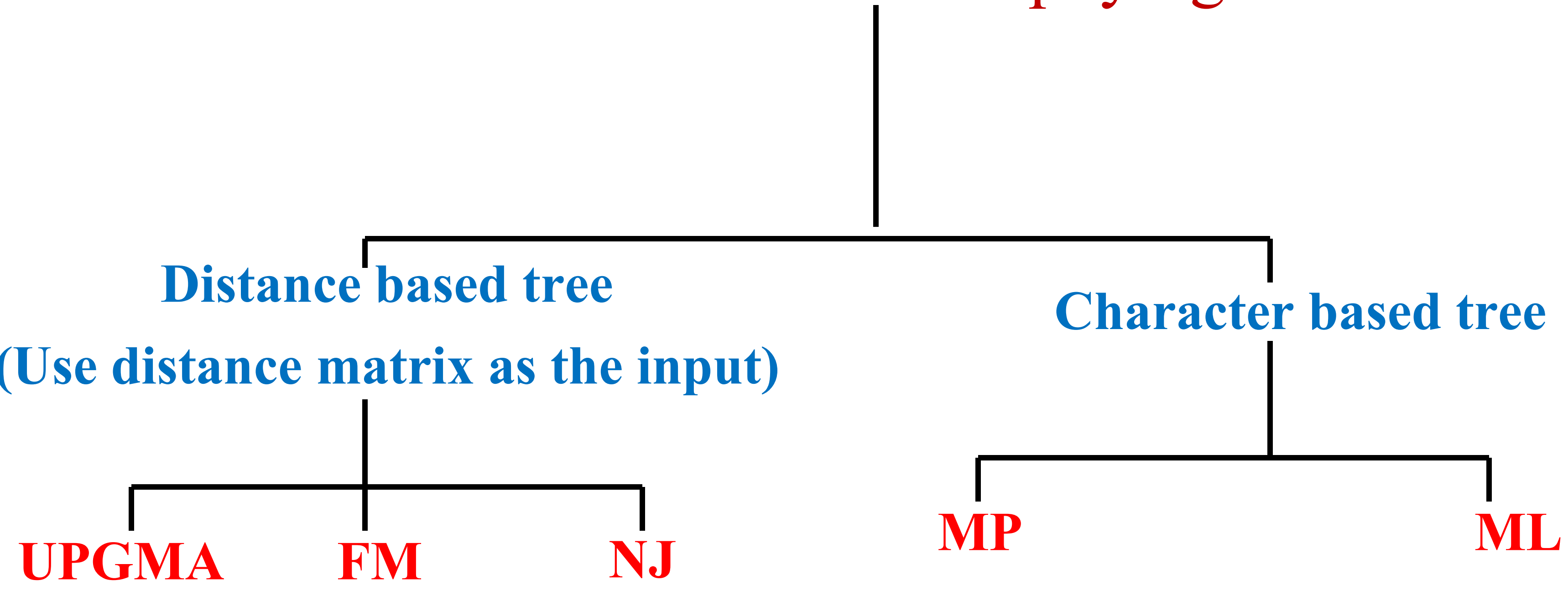
At some places, recognisable sequence similarity

No similarity, recognisable

Distance –based method

Maximum likelihood method

Methods for construction of phylogenetic tree



Distance based tree

(Use distance matrix as the input)

UPGMA

FM

NJ

Character based tree

MP

ML

UPGMA-Unweighted paired group with arithmetic mean method

FM-Fitch-Margoliash method

NJ-Neighbour joining method

ML-Maximum likelihood method

MP-Maximum parsimony method

Thank you