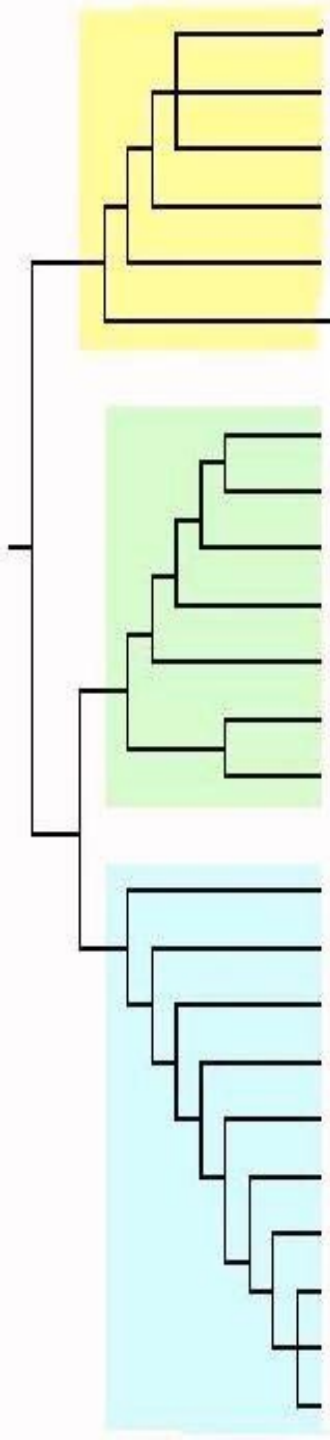
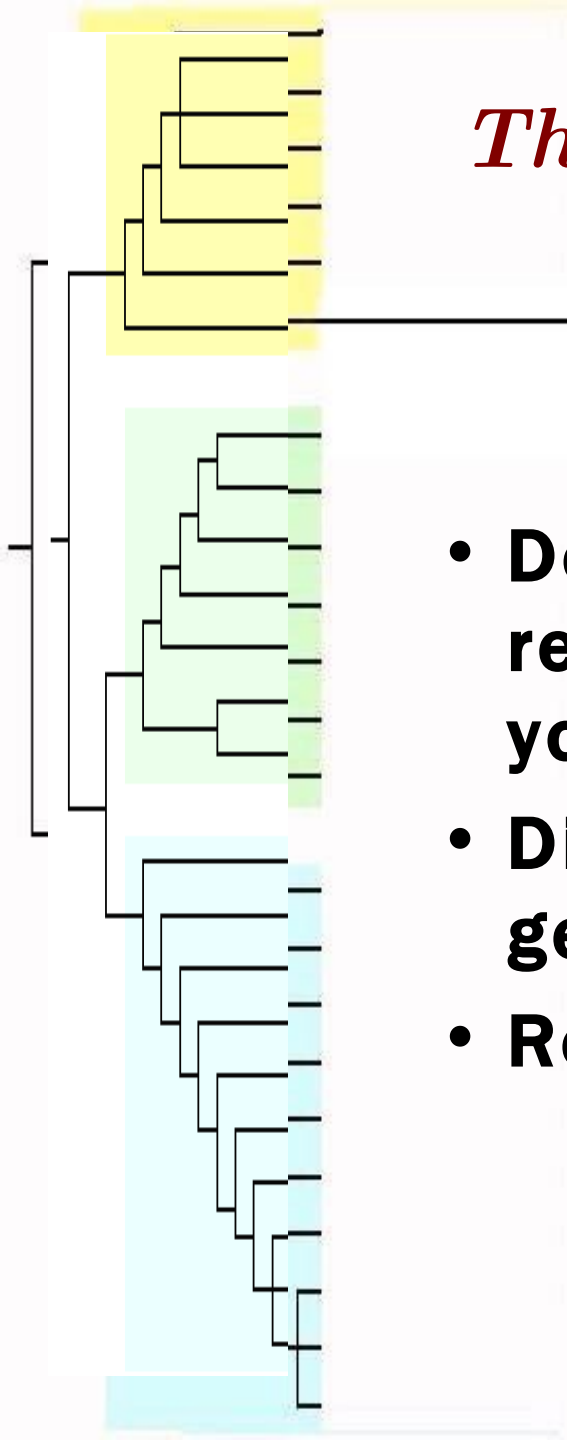


UNIT V

Softwares For Phylogentic Analysis

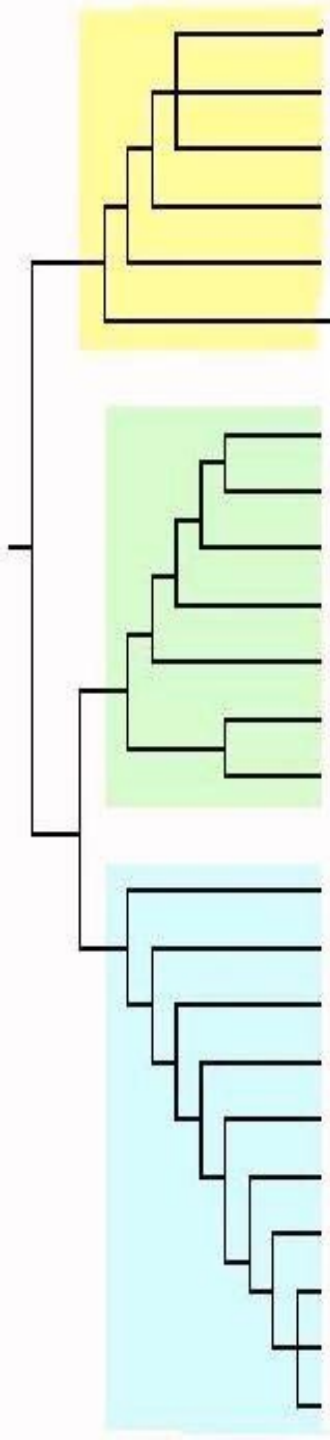
Survey of Software Programs Available For Phylogenetic Analysis

S. Prasanth Kumar, II M.Sc (BIOINFORMATICS), Alagappa University, India. prasanthbioinformatics@gmail.com

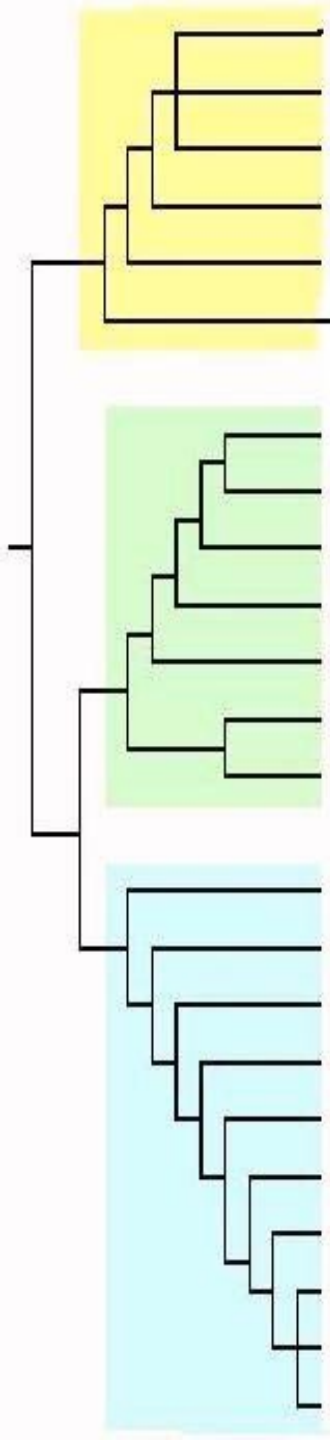


Three Major Reasons for Using Phylogenetics

- **Determining the closest relatives of the organism that you're interested**
- **Discovering the function of a gene**
- **Retracing the origin of a gene**



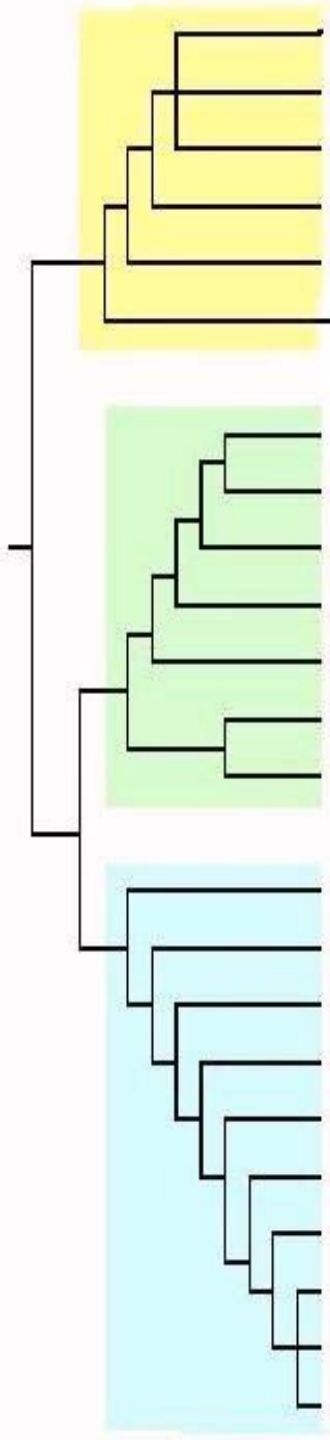
Survey of Various Phylogenetic Programs



Molecular Clock

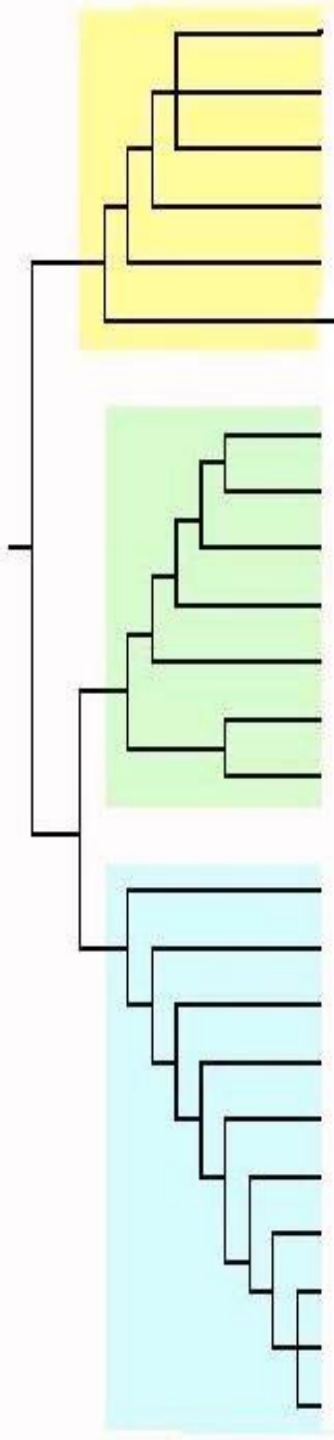
An assumption by which molecular sequences evolve at constant rates so that the amount of accumulated mutations is proportional to evolutionary time

Based on this hypothesis, branch lengths on a tree can be used to estimate divergence time



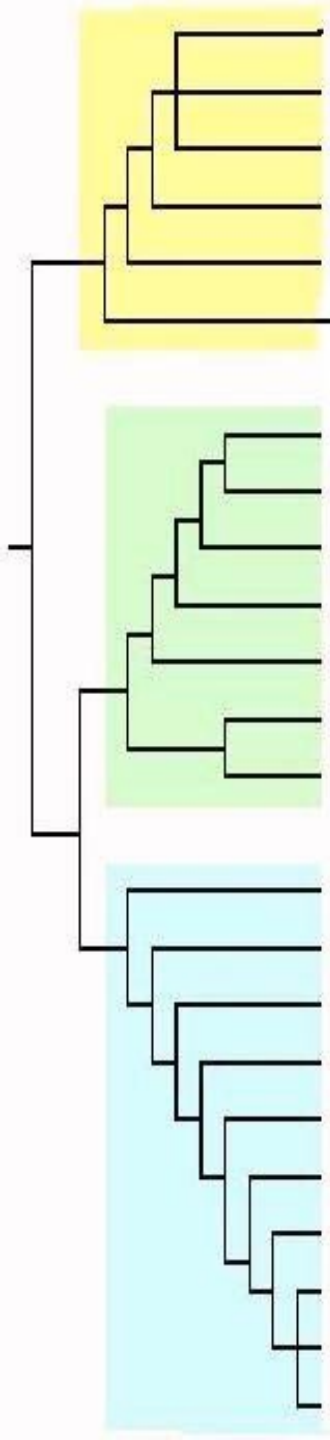
Desirable Qualities of Algorithms

- **Time consuming if large data sets are used for phylogenetic analysis**
- **Algorithms should be heuristic (i.e. a rule of thumb: skipping unnecessary steps and concentrating more on reliable steps)**
- **Recursion and Iteration steps for efficient tree construction**



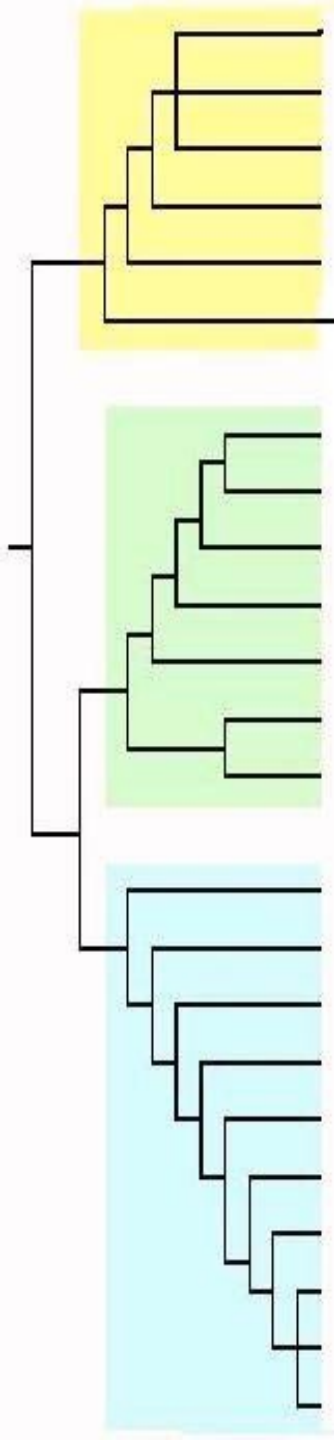
Substitution Models Assessment

- **All Substitution models to be used for correct phylogeny assessment**
- **Location of Transitions and Transversions e.g. Jukes-Cantor, Kimura model, etc**
- **Our phylogenetic analysis depends upon the MSA generated**
- **The MSA should concentrate on Structural alignments rather than Sequence alignments**



Desirable attributes of Phylogeny

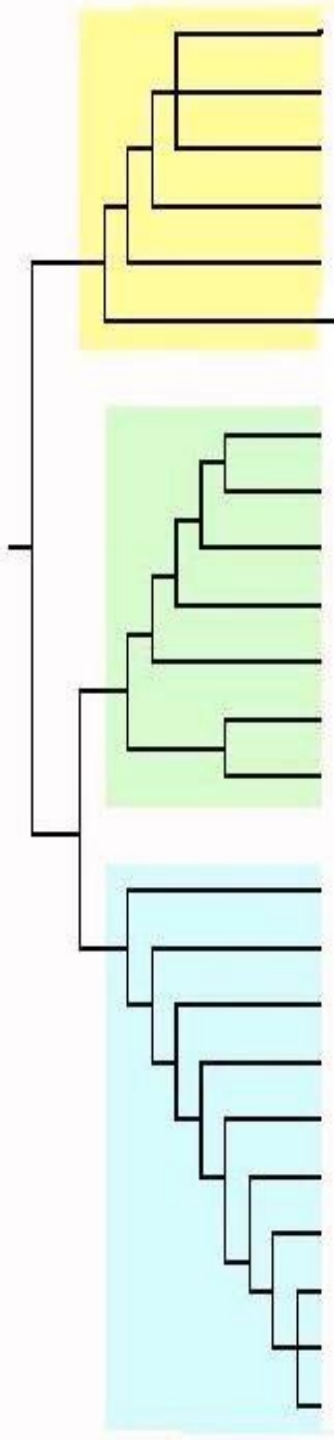
- **Manual inspection of MSA is always required. It can be done using MSA editors**
- **Methods used for phylogenetic analysis should contribute to all the following attributes:**
  - **Minimum Evolution**
  - **Evolutionary distance**
  - **Not purely based on Mathematical calculations**



Stress on "Initial Alignment"

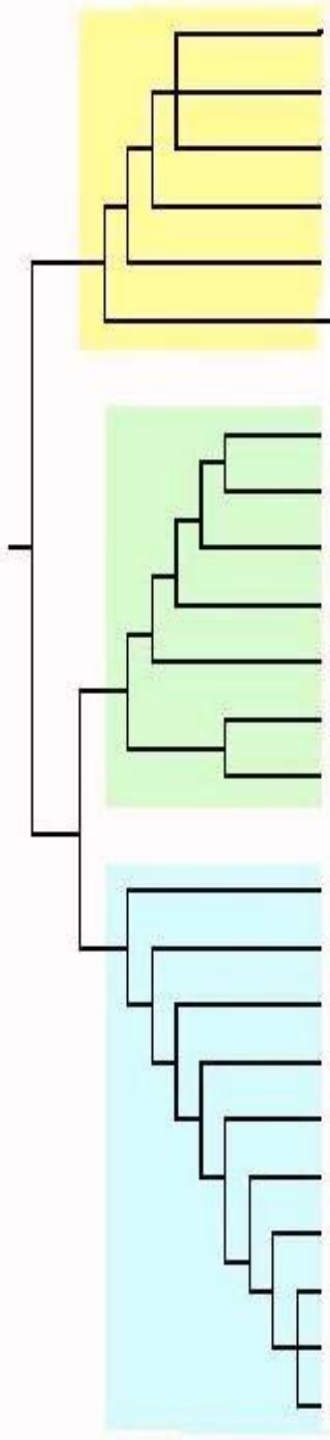
- **We should ensure that the phylogenetic tree has been constructed based on Informative sites (i.e. column of conserved residues)**

- **There should be no error in the initial alignment because it would affect the whole procedure to generate phylogenetic tree leading to erroneous tree (s)**



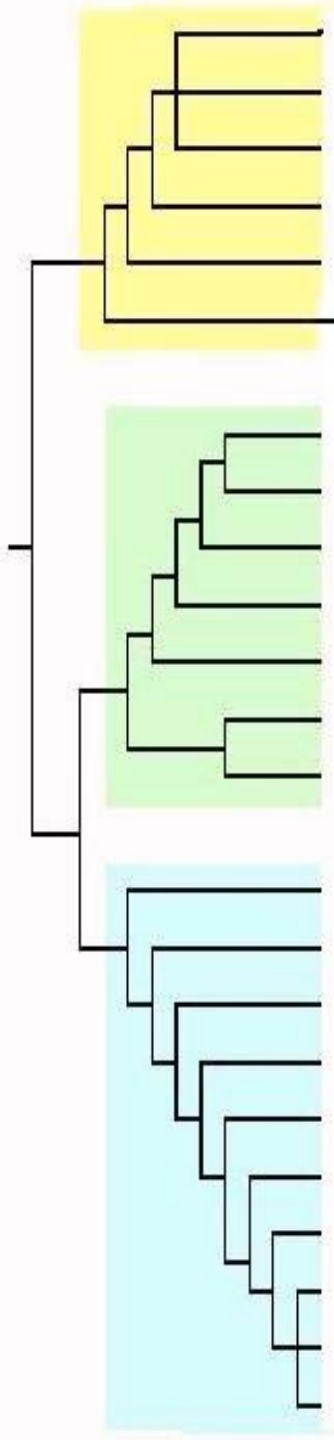
Choosing the Right Program

- **More than 1 substitution models is acceptable for efficient evolutionary predictions**
- **Randomness is accepted only in analyzing the reliability of the phylogenetic tree as in Bootstrapping technique**
- **Choosing the program should be dependent on our datasets (sequences)**



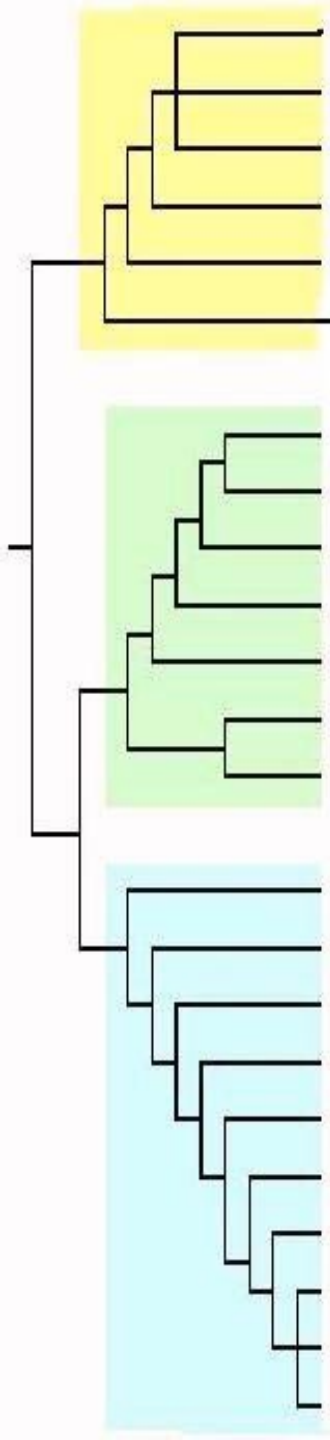
Different Topologies

- **If only one tree has been generated as does by Distance methods, Bootstrapping is appreciated**
- **It will be always good to work on different topologies of a phylogenetic tree**
- **Equal importance should be given to the sequences (excluding the pair of sequence taken for initial alignment) that has closest relationship**



Selecting Closest Sequences

- **Programs should consider Informative sites discarding Non informative sites and not on the alignment score of the informative sites**
- **Closest sequences does not mean biologically the pair of sequences having the highest alignment score, hence care should be given to choose molecular markers (explained in next slide)**



Choice of Molecular Markers

Different individuals within a population

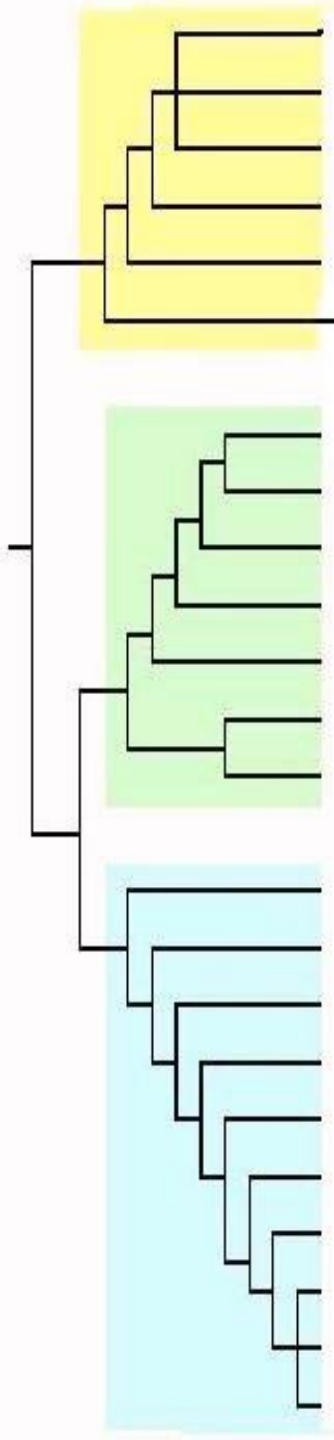
Non coding regions of mitochondrial DNA are often used.

More widely divergent groups of organisms

One may choose either slowly evolving nucleotide sequences, such as ribosomal RNA or protein sequences.

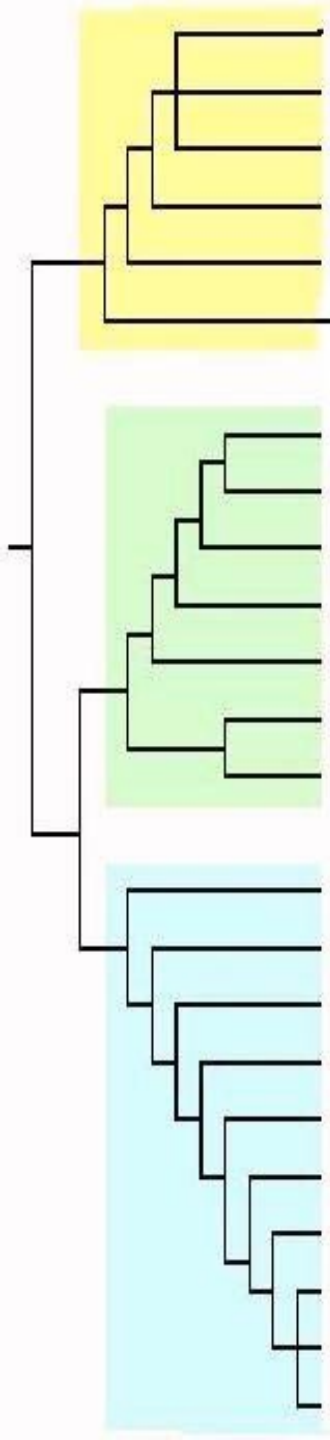
If the phylogenetic relationships to be delineated are at the deepest level, such as between bacteria and eukaryotes,

Conserved protein sequences makes more sense than using nucleotide sequences.



Codon Preference

- **Codon usage table for model organisms should be used if we are finding out phylogeny within a species using nucleotide species**
- **Methods exploring different topologies is appreciable since these methods explore minimum evolution**

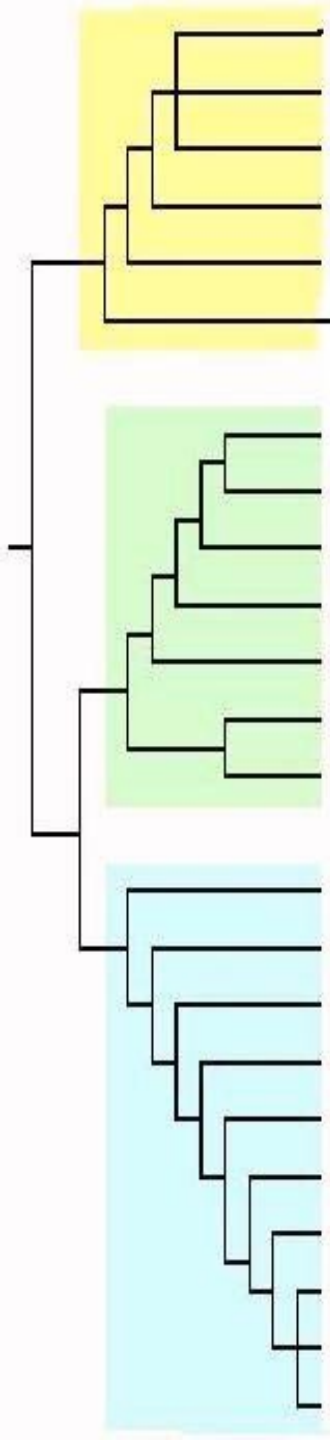


Weighted Parsimony

- **Weighted parsimony is better than Unweighted parsimony (treats all mutations as equivalent) mutations of some sites are known to occur less frequently than others,**

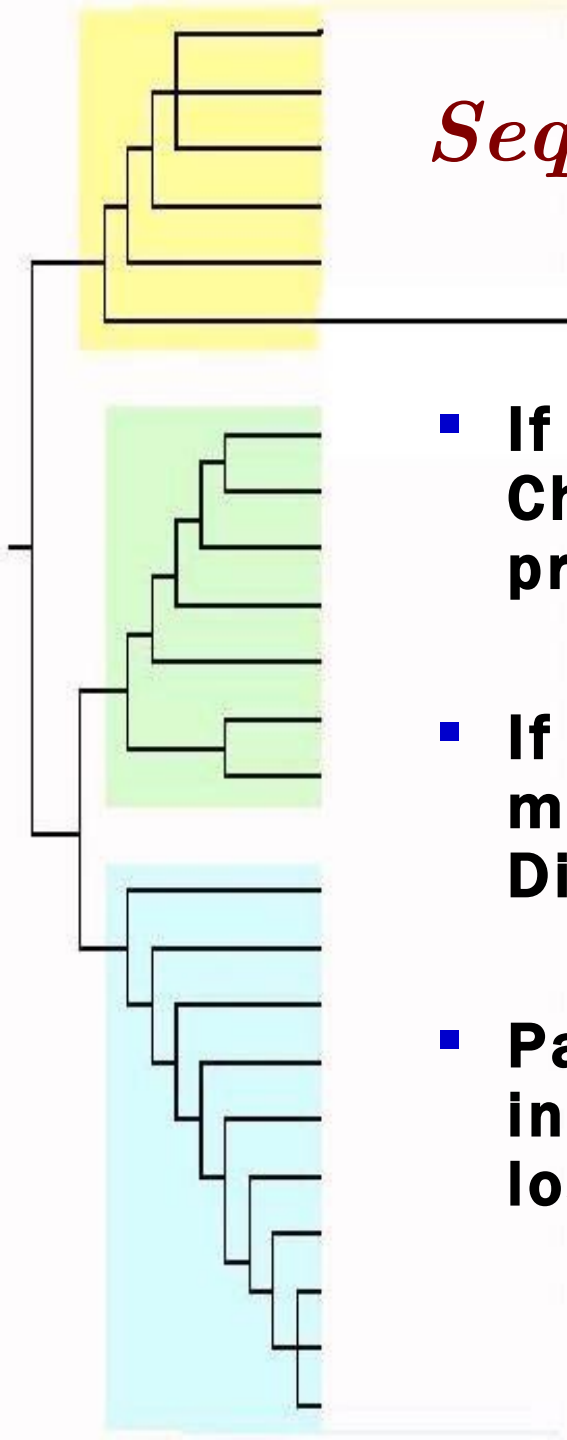
  for example, transversions versus transitions, functionally important sites versus neutral sites.

- **Therefore, a weighting scheme that takes into account the different kinds of mutations helps to select tree topologies more accurately**



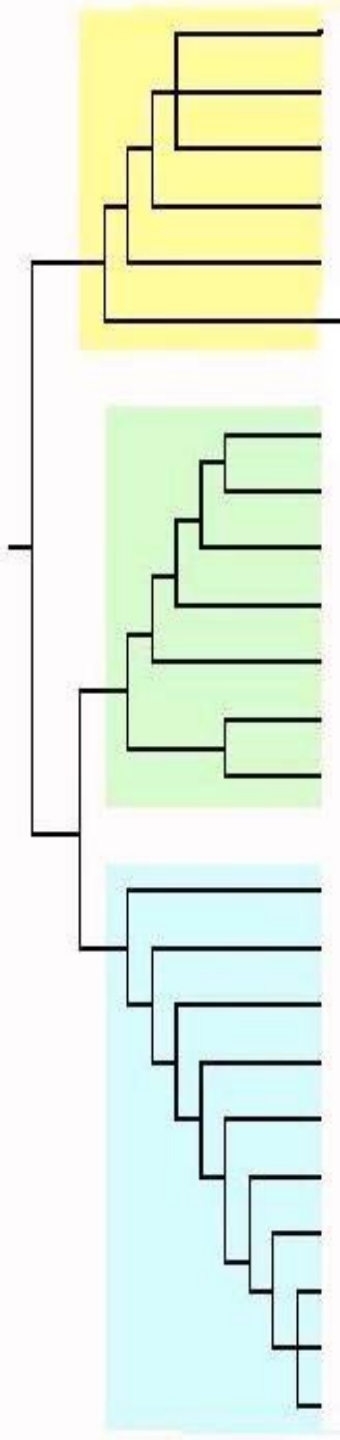
User Requirements

- The program should be user-friendly. It should display tree with higher graphical resolution
- Choosing substitution models for our study can be quite confusing. Hence, tools for selecting substitution models will be a good practice
- Modeltest, a program for selecting appropriate substitution models for nucleotide sequences



Sequence Divergence & Choice of Methods

- **If sequence divergence is low: go for Character based methods which can provide information about homoplasy**
- **If sequence divergence is high and the mount of homoplasies is large: go for Distance based methods**
- **Parsimonious tree considers only informative sites. There is chance of loss of phylogenetic signal**

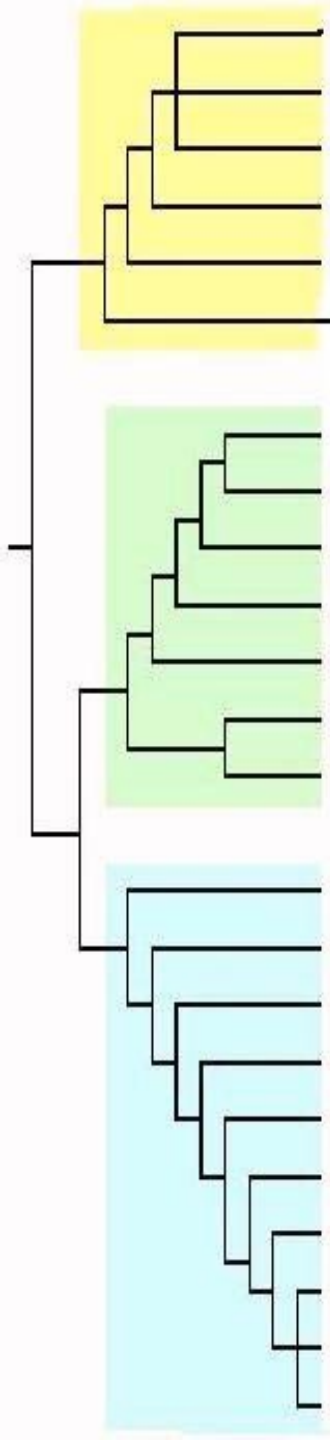


Tree Topology Assessment Heuristic Algorithms for Probability based Methods

- **Maximum Likelihood (ML) uses probabilistic models to choose a best tree that has the highest probability or likelihood of reproducing the observed data. Choosing an unrealistic substitution model may lead to an incorrect tree**

- **Some of the new heuristics used are:**
  - **Quartet Puzzling**
  - **Bayesian Inference**

Quartet Puzzling: the total number of taxa are divided into many subsets of four taxa known as *quartets*. An optimal ML tree is constructed from each of these quartets

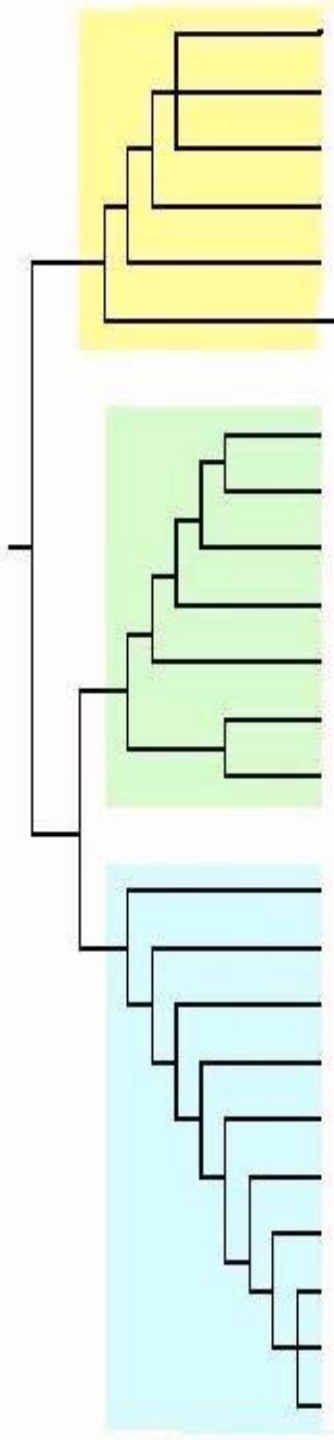


Tree Topology Assessment Heuristic Algorithms for Probability based Methods

Bayesian Analysis

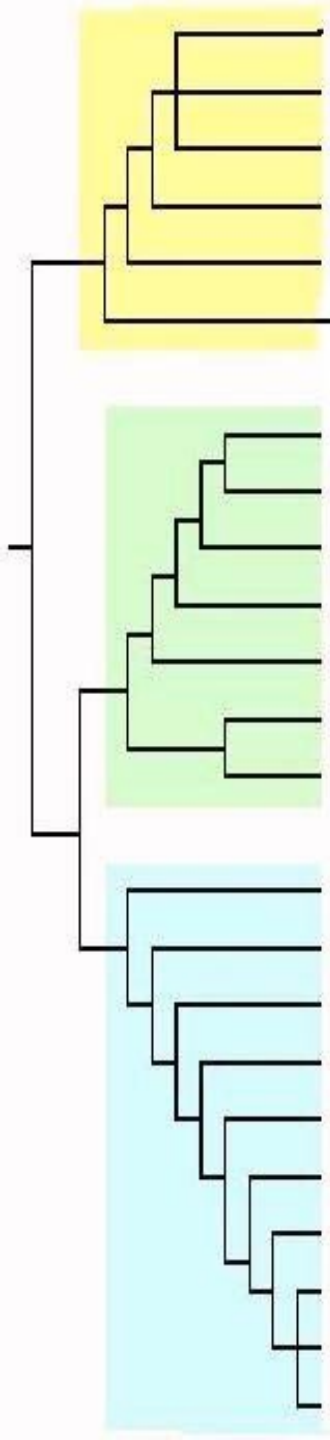
The essence of Bayesian analysis is to make inference on something unobserved based on existing observations. It makes use of an important concept of known as *posterior probability*, which is defined as the probability that is revised from prior expectations, after learning something new about the data

$$\text{Posterior probability} = \frac{\text{Prior probability} * \text{Conditional likelihood}}{\text{Total probability}}$$



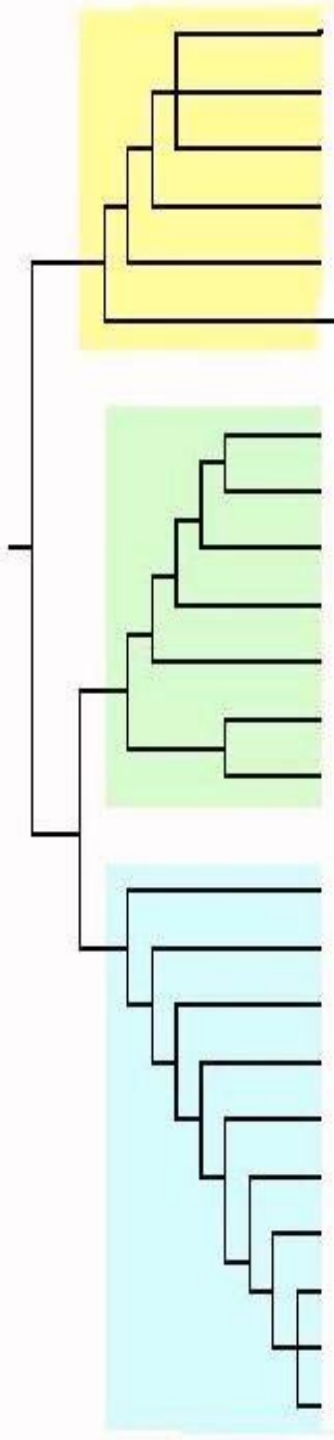
Analyzing Mutations

- **These mutations are easily detected by multiple alignment**
  - **Base Substitutions**
  - **Indel - Insertions & Deletions**
- **These mutations are not easily detected by multiple alignments**
  - **Transposition**
  - **Exon (domain) Shuffling**
- **Hence manual intervention is required. The organization of proteins should be addressed previously so as to identify these types of mutations**



Protein Sequences Are Preferable Than Nucleotide Sequences

- **Reason : Degeneracy of the genetic code**
- **“ 61 codons encode for 20 amino acids ”**
- **A change in a codon may not result in a change in amino acid**
- **DNA sequences are sometimes more biased than protein sequences because of preferential codon usage**



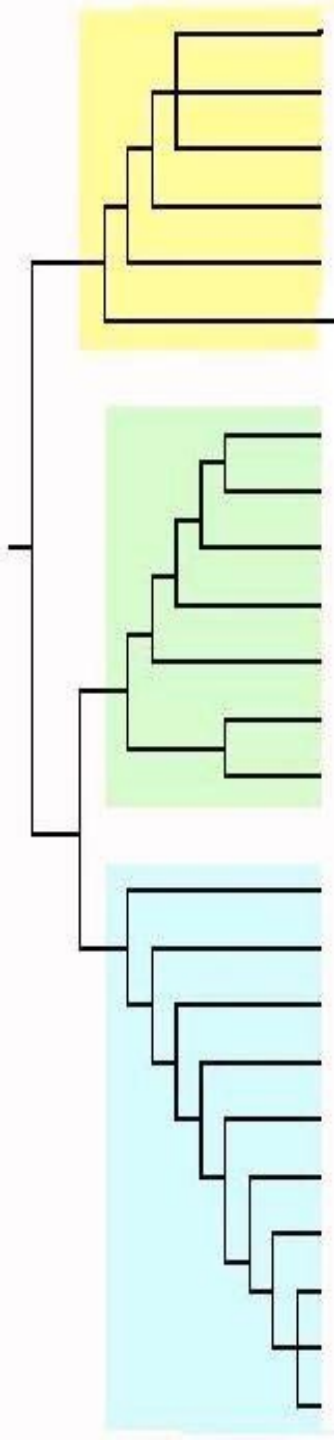
Improving Alignment Quality

Improve alignment by correcting alignment errors and removing potentially unrelated or highly divergent sequences

Programs: Rascal and NorMD

Detect and eliminate the poorly aligned positions and divergent regions

Programs: Gblocks



Deriving Ancestors

- **Strictly speaking, the identification or presence of ancestor in our datasets can be extinct.**

- **Adding a root (Ancestor) is often be desirable. To derive the ancestor, OUTGROUP (a sequence which has less divergence to our sequences, from which the evolutionary distance will be calculated, this sequence will not be for phylogentic tree construction) is selected**

Softwares available in WWW

NAME	Description	Methods
BEAST	Bayesian Evolutionary Analysis Sampling Trees	Bayesian inference, relaxed molecular clock, demographic history
Bosque	Integrated graphical software to perform phylogenetic analyses, from the importing of sequences to the plotting and graphical edition of trees and alignments	Distance and maximum likelihood methods (through phyml, phylip & tree-puzzle)
ClustalW	Progressive multiple sequence alignment	Distance matrix/nearest neighbor
fastDNAml	Optimized maximum likelihood (nucleotides only)	Maximum likelihood
Geneious	Geneious provides sophisticated genome and proteome research tools	Neighbor-joining, UPGMA, MrBayes plugin, PHYML plugin

Softwares available in WWW

NAME	Description	Methods
HyPhy	Hypothesis testing using phylogenies	Maximum likelihood, neighbor-joining, clustering techniques, distance matrices
MEGA	Molecular Evolutionary Genetics Analysis	Distance, Parsimony and Maximum Composite Likelihood Methods
MOLPHY	Molecular phylogenetics (protein or nucleotide	Maximum likelihood
MrBayes	Posterior probability estimation	Bayesian inference
PAML	Phylogenetic analysis by maximum likelihood	Maximum likelihood
PAUP	Phylogenetic analysis using parsimony	Maximum parsimony, distance matrix, maximum likelihood
PHYLIP	Phylogenetic inference package	Maximum parsimony, distance matrix, maximum likelihood

Softwares available in WWW

NAME	Description	Methods
PhyloQuart	Quartet implementation (uses sequences or distances	Quartet method
TreeGen	Tree construction given precomputed distance data	Distance matrix
TREE-PUZZLE	Maximum likelihood and statistical analysis	Maximum likelihood
SplitsTree	Tree and network program	Computation, visualization and exploration of phylogenetic trees and networks