# Molecular Evolution

"Phylogenetic reconstruction is a fast-growing field that is enriched by different statistical approaches and by findings and applications in a broad range of biological areas. Fundamental to these are the mathematical models used to describe the patterns of DNA base substitution and amino acid replacement. These may become some of the basic models for comparative genome research."

—Pietro Liò and Nick Goldman [LIO1998].

## Introduction

The comparison between genome sequences of different organisms of the same species or from a different species show that these are not static but change or mutate over their evolutionary history. Changes or mutations in a genome can occur due to a variety of causes such as errors in DNA replication or external factors like UV rays. These changes or mutations may be **neutral, defective or advantageous** for adaptation, survival and reproduction of the relevant body. If a mutation occurs in the **germline** of an organism can be transmitted to their descendants. Thus a mutation that is neutral or advantageous for the reproduction of an organism can spread in a population and resulting fixed polymorphisms. A **polymorphism** is the existence of different variants of a DNA sequence called **alleles**.
The most frequent polymorphisms are **SNPs**, single nucleotide polymorphisms. These variations consist of a single nucleotide change by another between alleles. The **STR**, short tandem repeats, are the second most frequent polymorphisms. This consists of repeating a different number of times in different alleles of short sequences of DNA. Finally, **indels, transpositions, inversions and duplications** may appear as polymorphisms rarely.

## Molecular Phylogeny

The phylogeny is the branch of biology focused on the inference of **evolutionary relationships** between existing species. Traditionally, the study was based on morphological and physiological character. In early 1980 with the acquisition of the first gene and protein sequences raises the molecular phylogeny. This discipline is based on a comparison of biological sequences to perform a **hierarchical classification** between existing species.
The main objectives of the molecular phylogeny are:

- Determine a hierarchical relationship between existing species according to their evolutionary relationship. So that organisms that share a common ancestor are grouped close earlier than those with a more distant common ancestor.
- Estimating the divergence time between species, i.e., the time of existence of the nearest common ancestor.

## Phylogenetic Analysis

The determination of the hierarchical relationship between species and divergence time estimation among them based on biological sequences such as proteins or DNA is performed by **phylogenetic studies**. Phylogenetic studies typically are divided into five distinct phases:

- **Phase 1:** Selecting of the biological sequences to be analyzed.
- **Phase 2:** Building a multiple alignment of the biological sequences selected.
- **Phase 3:** Selecting the substitution models, statistical models of molecular evolution, for the corresponding sequences.
- **Phase 4:** Building the phylogenetic trees based on the corresponding multiple alignments and substitution models.
- **Phase 5:** Statistical evaluation of the phylogenetic trees.

These phases are not organized in a linear fashion. It is common to have to go back to previous phases to review some decisions made before advancing to the next phase. It is also frequent complete the study for various possible choices in the different phases and compare

results before making a final decision on the phylogenetic analysis performed. For example, normally

will be taken several multiple alignments in phase 2 with different parameters, to explore different substitution models in phase 3 and to build phylogenetic trees with several different methods in phase 4. Different results will be independently evaluated in phase 5 and compared together for, based on existing literature and biological information available, making a final decision on the most informative analysis.

## Phylogenetic Analysis

### Selecting the biological sequences

Phylogenetic studies are based on the comparison of homologous biological sequences. The study of the differences between them allows to estimate the evolutionary relationship between the corresponding species and their divergence time.

To obtain homologous sequences, databases such as RefSeq, Uniprot or HomoloGene can be very useful. Another way to obtain homologous sequences is through alignments using BLAST [1] results.

### Multiple alignment of the biological sequences

The multiple alignment of the analyzed biological sequences is the most important step in phylogenetic studies as it involves the comparison between the different sequences.

Is necessary to check the following points:

- Remove non-homologous sequences, those that show no alignment. There must be a pair sequence alignment and assess their significance.
- If alignment is not good and it is certain of homology between sequences, the parameters of insertion penalty and gap extension must be modified.
- Normally there is no known of the whole sequence of those corresponding sequences and there are many gaps. Is necessary to eliminate the columns corresponding to gap.
- Duality between multiple alignment and phylogenetic tree.

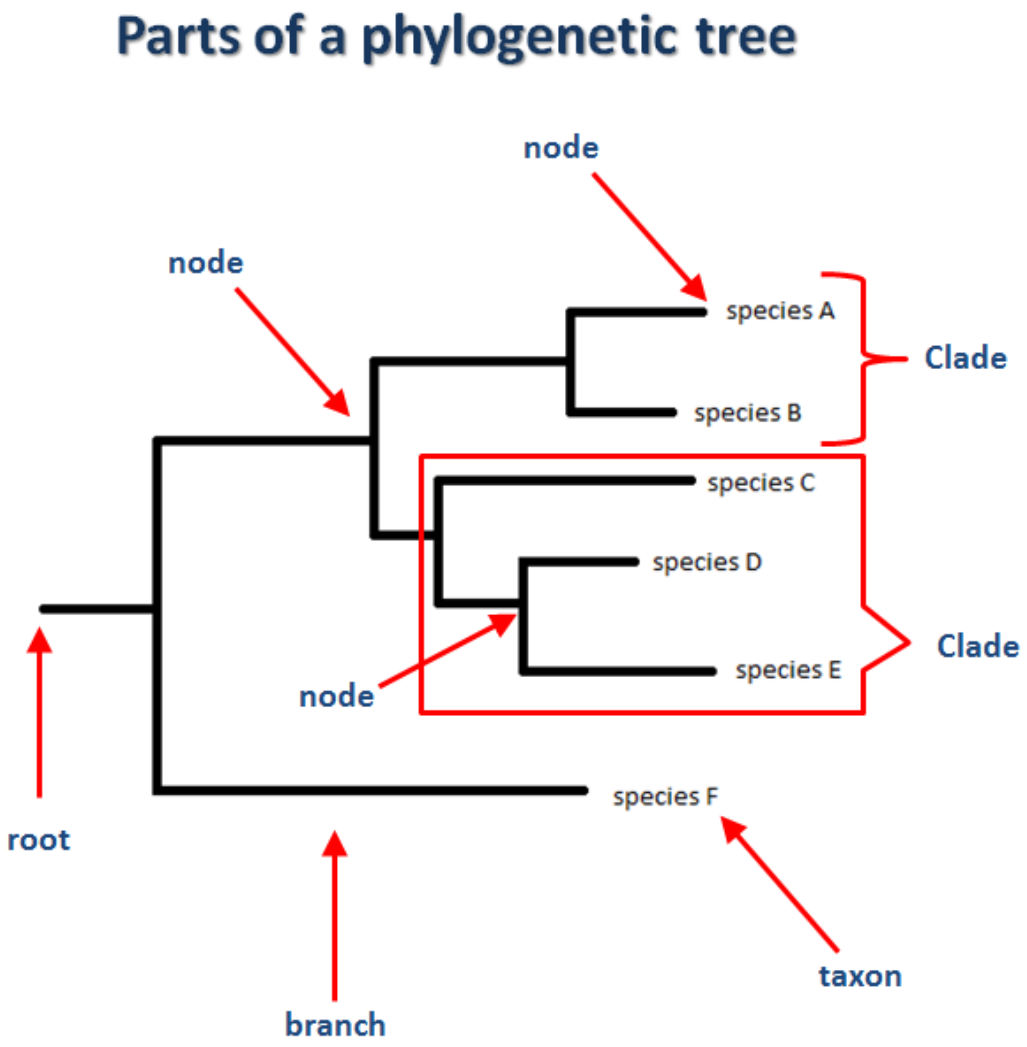
There are several tools to perform multiple sequence alignment, MEGA [2] is one of them. This tool allows to align multiple sequence via MUSCLE [3] and ClustalW [4] alignment methods

### Selecting the substitution models

After completing the multiple alignment, is possible to estimate the **genetic distance** between the different sequences. The genetic distance between two homologous sequences is defined as the number of substitutions accumulated between them since they diverged from a common ancestor. Estimation of genetic distance is not trivial since **not all substitutions are observable** especially in sequences with many substitutions.
Figure 6.1 shows why not all substitutions are observable, there are actual substitutions that cannot be observed.

### Building phylogenetic trees

The main objects of study in molecular phylogeny: the establishment of hierarchical relationships between species according to their evolutionary relationship and the estimated time of divergence between species are represented by using phylogenetic trees.




Parts of a phylogenetic tree

There are two types of trees according to the existence of an outstanding node called **root**: **Rooted trees** have a node called root, which corresponds to the common ancestor of all the taxa. In rooted trees can be establish a relationship of temporality. On the other hand **unrooted trees** lack of a root node, for that reason can not establish a temporal relationship.

There are mainly two methods to determine the root of an unrooted tree:

- Added an **outgroup**, a taxon that is known to be the farthest from the rest, and determine the root at the midpoint of the branch which joins the clade composed of the remaining taxa.
- Determine the longest branch and set the root at its midpoint.

**Tree-Building Methods**

The most popular and frequently used methods of tree building can be classified into two major categories: **phenetic methods based on distances** and **cladistic methods based on**

**characters**. The former measures the pair-wise distance/dissimilarity between two genes, the actual size of which depends on different definitions, and constructs the tree totally from the resultant distance matrix. The latter evaluate all possible trees and seek for the one that optimizes the evolution.

*Distance-Based Methods:*
The most popular distance-based methods are the unweighted pair group method with arithmetic mean (UPGMA), neighbor joining (NJ) and those that optimize the additivity of a distance tree (FM and ME).

- **UPGMA Method**: This method follows a clustering procedure:
  1. Assume that initially each species is a cluster on its own.
  2. Join closest 2 clusters and recalculate distance of the joint pair by taking the average.
  3. Repeat this process until all species are connected in a single cluster.

Strictly speaking, this algorithm is phenetic, which does not aim to reflect evolutionary descent. It assigns equal weight on the distance and assumes a randomized molecular clock. WPGMA is a similar algorithm but assigns different weight on the distances.

UPGMS method is simple, fast and has been extensively used in literature. However, it behaves poorly at most cases where the above presumptions are not met.

- **Neighbor Joining Method (NJ)**: This algorithm does not make the assumption of molecular clock and adjust for the rate variation among branches. It begins with an unresolved star-like tree. Each pair is evaluated for being joined and the sum of all branches length is calculated of the resultant tree. The pair that yields the smallest sum is considered the closest neighbors and is thus joined. A new branch is inserted between them and the rest of the tree and the branch length is recalculated. This process is repeated until only one terminal is present.

NJ method is comparatively rapid and generally gives better results than UPGMA method. But it produces only one tree and neglects other possible trees, which might be as good as NJ trees, if not significantly better. Moreover, since errors in distance estimates are exponentially larger for longer distances, under some condition, this method will yield a biased tree.

- **Weighted Neighbor-Joining (Weighbor)**: The Weighbor criterion consists of two terms; an additivity term (of external branches) and a positivity term (of internal branches), that quantifies the implications of joining the pair. Weighbor gives less weight to the longer distances in the distance matrix and the resulting trees are less sensitive to specific biases than NJ and relatively immune to the "long branches attraction/distraction" drawbacks observed with other methods.
- **Fitch-Margoliash (FM) and Minimum Evolution (ME) Methods**: Fitch and Margoliash proposed in 1967 a criteria (FM Method) for fitting trees to distance matrices. This method seeks the least squared fit of all observed pair-wise distances to the expected distance of a tree. The ME method also seeks the tree with the minimum sum of branch lengths. But instead of using all the pair-wise distances as FM, it fixed the internal nodes by using the distance to external nodes and then optimizes the internal branch lengths.

FM and ME methods perform best in the group of distance-based methods, but they work much more slowly than NJ, which generally yield a very close tree to these methods.

*Character-Based Methods:*
Distance-based methods are more rapid and less computationally intensive than character-based methods, but the actual characters are discarded once the distance matrix is derived. On the other hand, character-based methods make use of all known evolutionary information, i.e.

the individual substitutions among the sequences, to determine the most likely ancestral relationships.

- **Maximum parsimony (MP)**: The criterion of MP method is that the simplest explanation of the data is preferred, because it requires the fewest conjectures. By this criterion, the MP tree is the one with fewest substitutions/evolutionary changes for all sequences to derive from a common ancestor.

For each site in the alignment, all possible trees are evaluated and are given a score based on the number of evolutionary changes needed to produce the observed sequence changes. The best tree is thus the one that minimized the overall number of mutation at all site.

MP works faster than ML and the weighted parsimony schemes can deal with most of the different models used by ML. However, this method yields little information about the branch lengths and suffers badly from long-branch attraction, that is the long branches have become artificially connected because of accumulation of inhomogous similarities, even if they are not at all phylogenetically related.

- **Maximum Likelihood (ML)**: Like MP methods, ML method also uses each position in an alignment and evaluates all possible trees. It calculates the likelihood for each tree and seeks the one with the maximum likelihood.

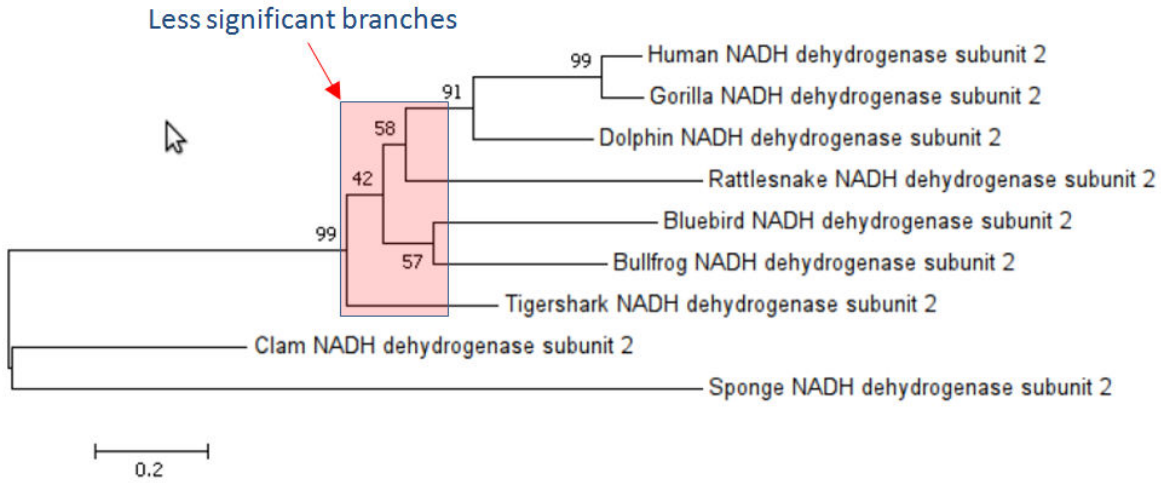
For a given tree, at each site, the likelihood is determined by evaluating the probability that a certain evolutionary model has generated the observed data. The likelihood's for each site are then multiplied to provide likelihood for each tree.

ML method is the slowest and most computationally intensive method, though it seems to give the best result and the most informative tree.

### Statistical evaluation of phylogenetic trees

Once constructed a phylogenetic tree should be evaluated its robustness [HOLMES2002], which means being able to answer the following question: How often does is obtained a given branching order considering similar sequences to those used?
The assessment trees method most widespread consists of Bootstrapping [HOLMES2003]. In this method similar sequences to those used are constructed as permutations with repetition of the multiple alignment of the corresponding columns. A new tree is constructed with the new alignment and this process is repeated a certain number of times.
To each branch is assigned a percentage of occurrences in the constructed trees and in this way is assumed that a branch is significant if it appears more than 50% or 70% of the times. Remaining branches may be condensed resulting in polytomy. It is usually to build also the consensus tree for including the most frequent branching order.

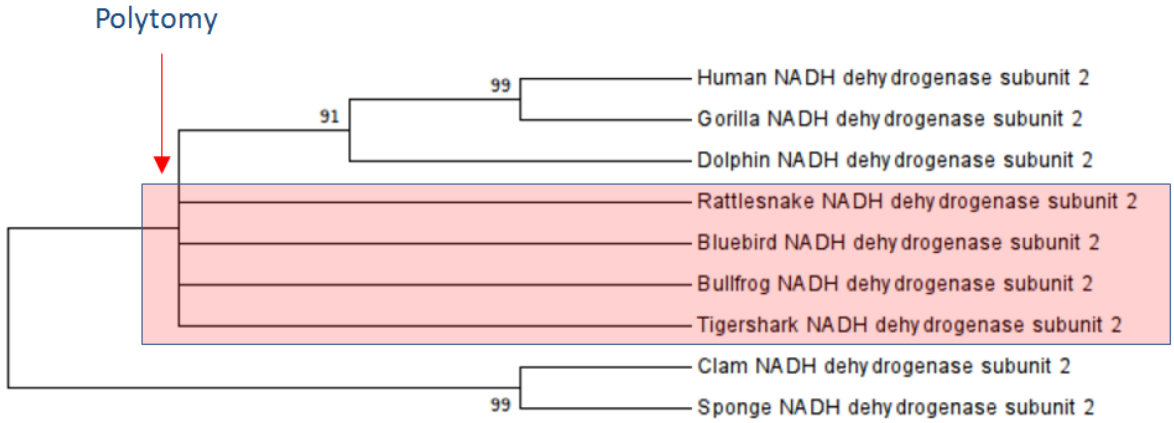
## Bootstrap consensus tree




Bootstrap consensus tree.

This Figure shows that not all branches are significant, therefore is necessary to build a condensed tree to give more significance to those branches

## Bootstrap condensed tree




Bootstrap condensed tree.

This Figure shows how the less significant branches have been changed into polytomy.

**References**


1. Basic Local Alignment Search Tool (BLAST) is the tool most frequently used for calculating sequence similarity. BLAST comes in variations for use with different query sequences against different databases.http://blast.ncbi.nlm.nih.gov
2. Molecular Evolutionary Genetics Analysis (MEGA) is an integrated tool for conducting automatic and manual sequence alignment, inferring phylogenetic trees, mining web-based databases, estimating rates of molecular evolution, inferring ancestral sequences, and testing evolutionary hypotheses.



Prof. Rajarshi Kumar Gaur, Department of Biotechnology, D.D.U. Gorakhpur University
E-mail: gaurrajarshi@hotmail.com/rajarshi.biotech@ddugu.ac.in

3. Multiple Sequence Comparison by Log-Expectation (MUSCLE) is a program for creating multiple alignments of amino acid or nucleotide sequences.
4. Clustal W is a general-purpose multiple alignment program for DNA or proteins.