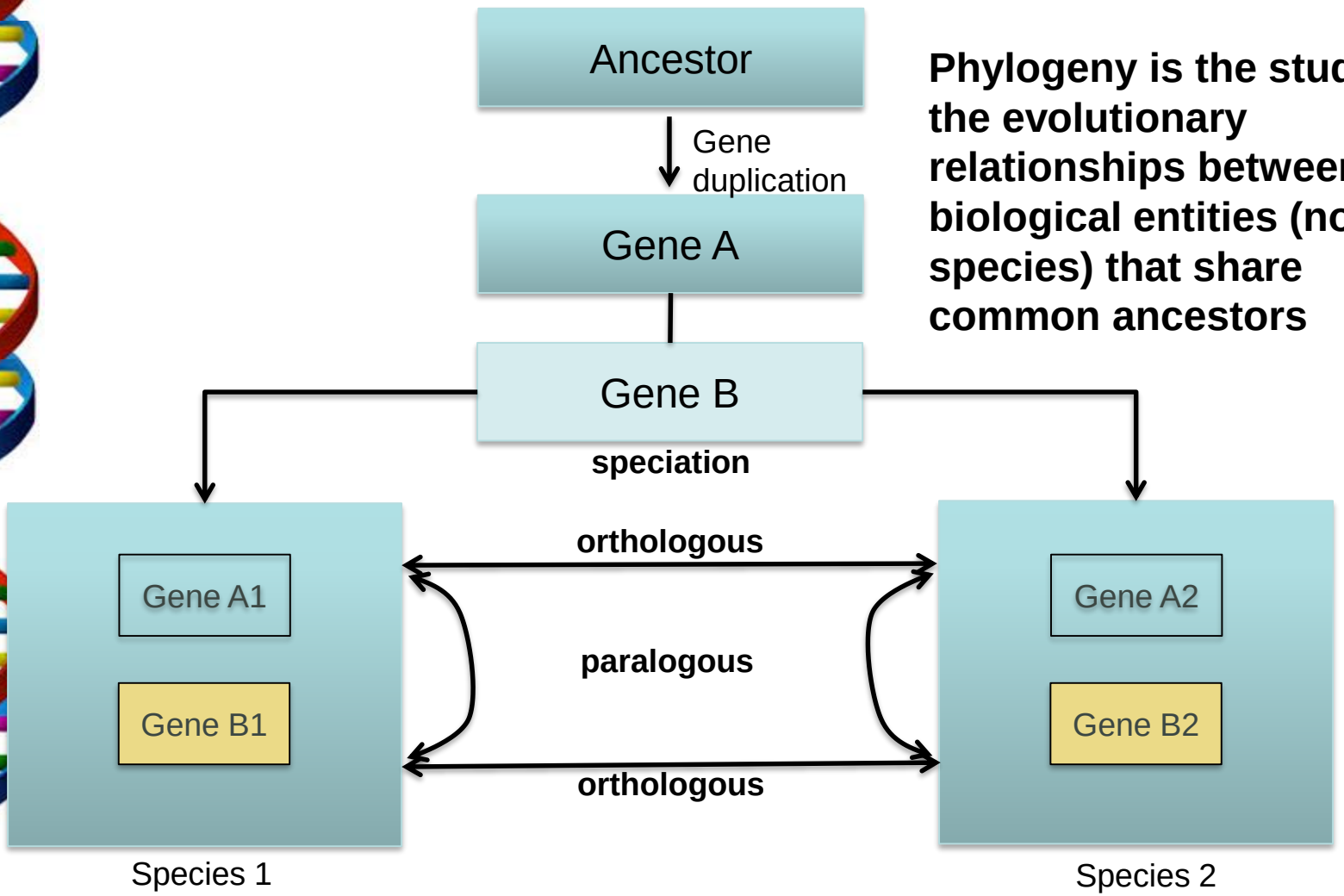
Phylogenetic Analysis

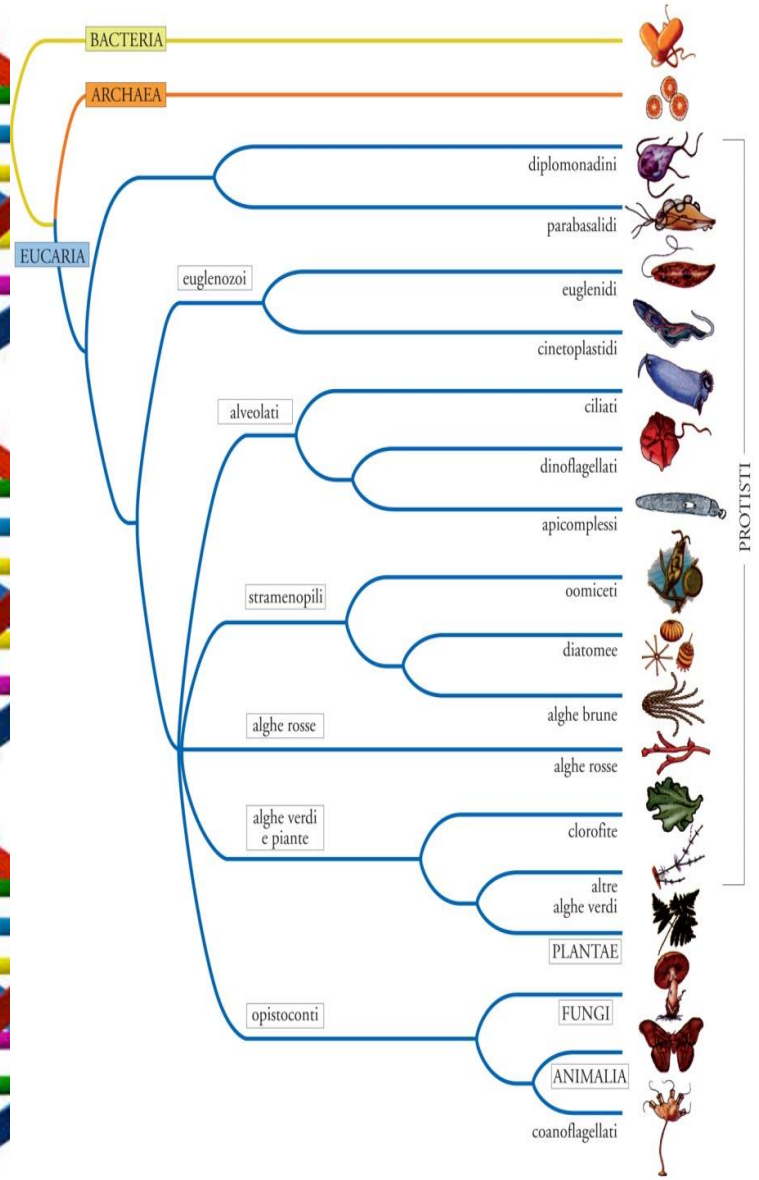
Dr. Antinisca DI MARCO
antinisca.dimarco@univaq.it

Phylogenetic Analysis



Phylogeny is the study of the evolutionary relationships between biological entities (not only species) that share common ancestors

Phylogenetic Analysis



Its graphical representation is the phylogenetic tree

The phylogenetic tree shows the time and temporal patterns of divergence processes

All organisms have a unique common ancestor in the past

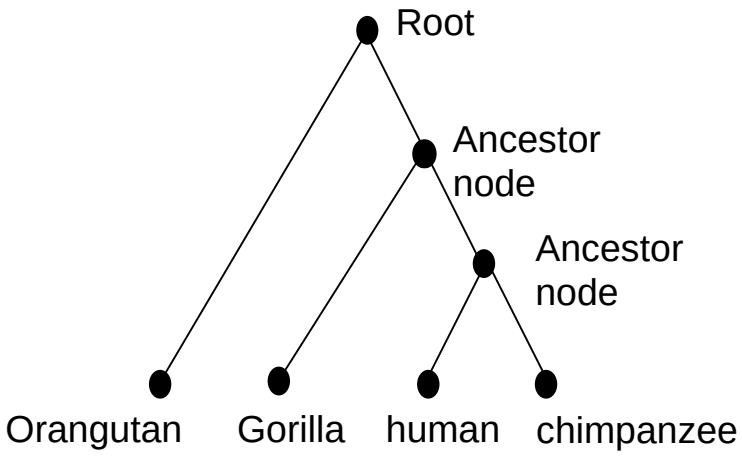
Each pair of organisms have a common ancestor in the past

speciation events follow each other in time by creating new species

Phylogenetic Analysis

Reconstruction of Phylogeny

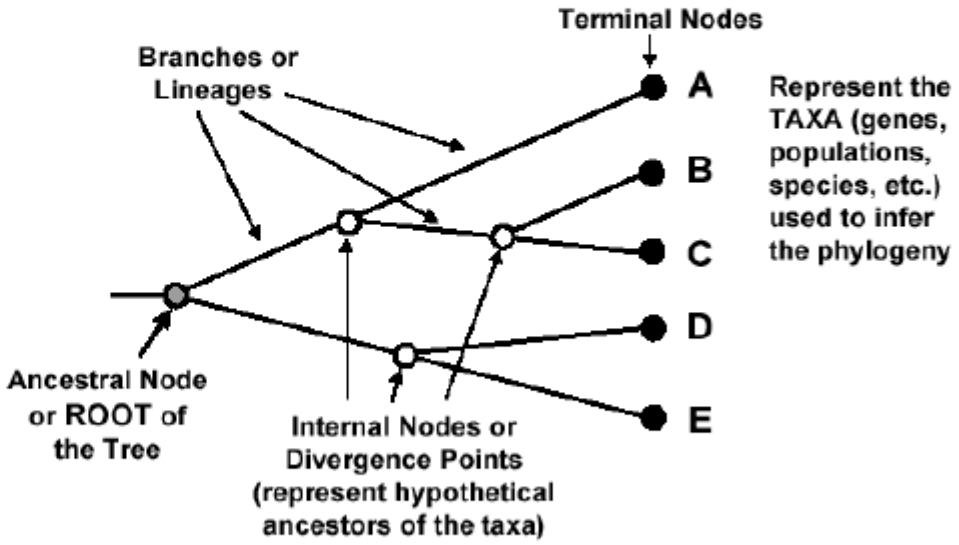
- An evolutionary tree, or phylogeny, is a tree with or without root whose internal nodes have at least grade 3 (with the exception of the root that has degree 2) and represent progenitor species, while the leaves represent existing species.
- The arcs of the tree usually represent the temporal distance between two species (nodes).



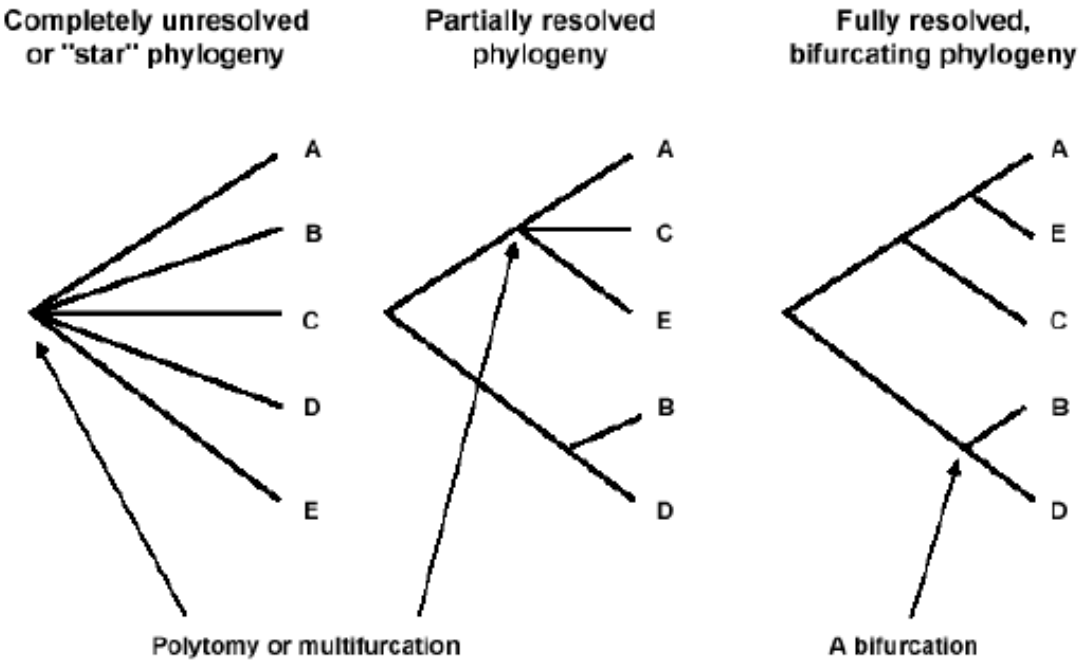
Terminology of phylogenetic trees

A tree is composed by:

- Terminal nodes o taxa that represents existing objects;
- Intermediate nodes o divergent pointing o branch that represents putative taxa ancestors
- An initial node (only in case of rooted tree) o root that represents the ancestor of all taxa.
- Edges or lines that link the different nodes

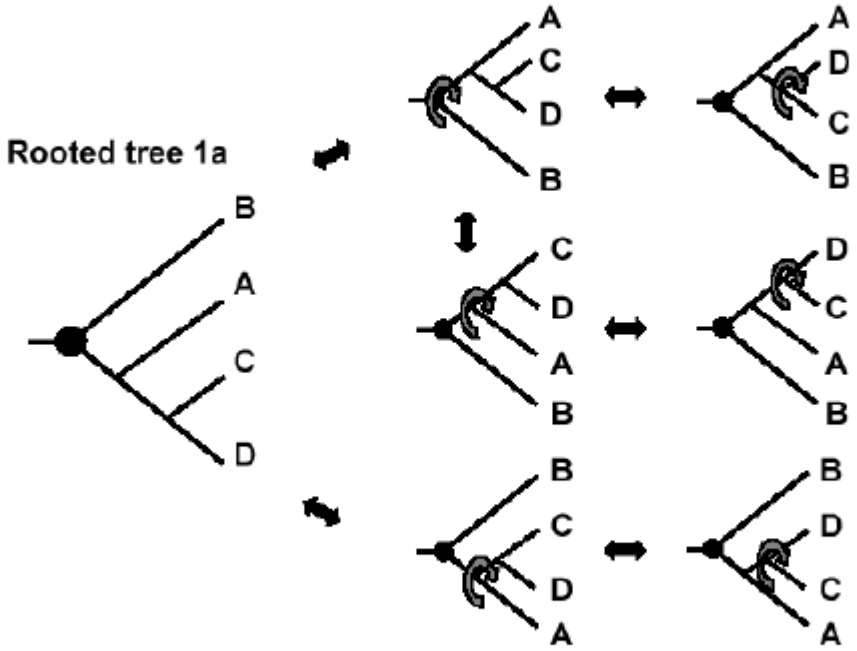


Determination of the branching taxa order



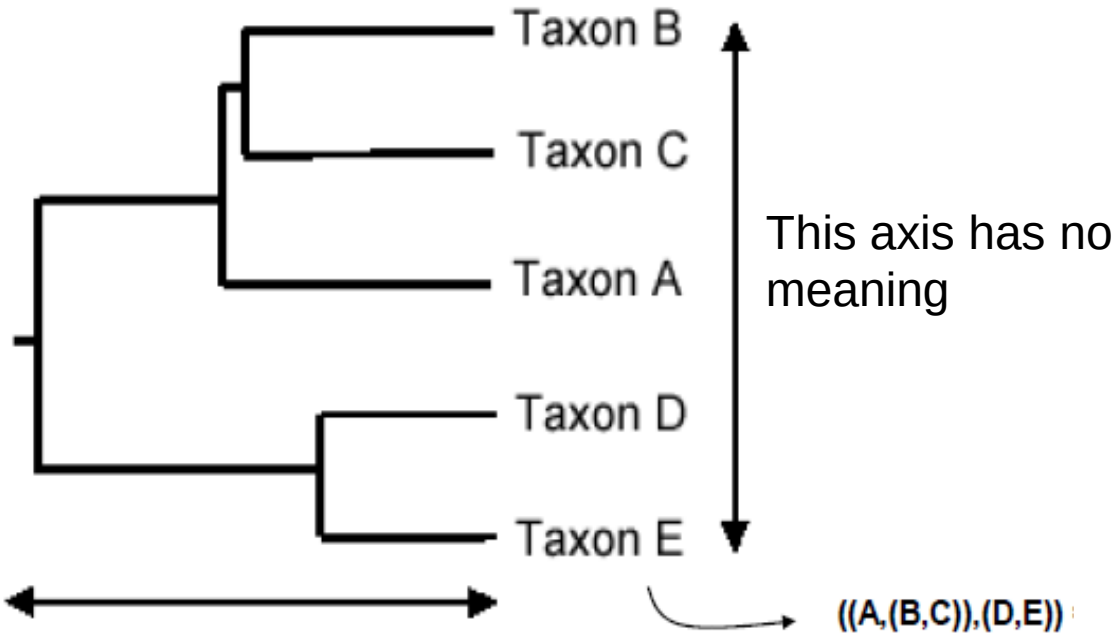
The aim of the inference of the phylogenetics is the determination of the taxa branching order. This corresponds to a tree that presents only bifurcations (dichotomy). A multifurcation (politomia) can be *soft*, that is resolvable with the addition of more phylogenetic data, or *hard,* i.e., caused by taxa simultaneous separation

The rotation of a node does not change the tree topology



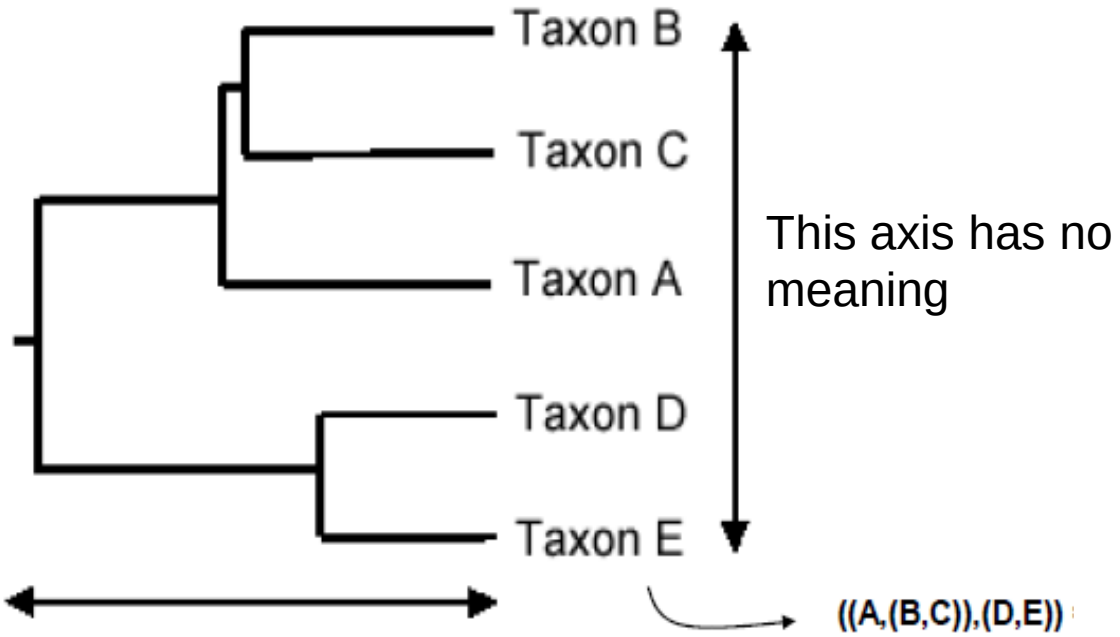
All rotations around a node provide trees with equivalent topology

representation of the evolutionary relationships



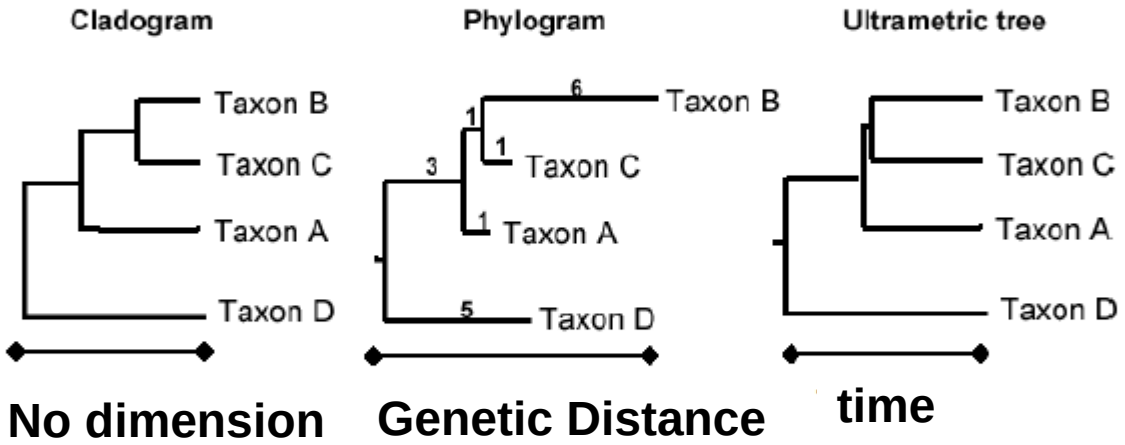
this axis can have any scale (cladogram) or be proportional to genetic distance (filogramma, or additives trees) or be proportional to the time (ultrametric trees).

representation of the evolutionary relationships



The tree and the expression with parentheses represent the same evolutionary relationships. For instance, B and C are closer to each other than A to each of the two. A, B, C form a clade. It is the sister group of the clade consisting of D and E. In a tree with a time scale, D and E are also the closest ones in the tree.

Three different types of trees



these trees have the same topology that represents the same evolutionary relationships among taza. The meaning of the arms is different in all three cases

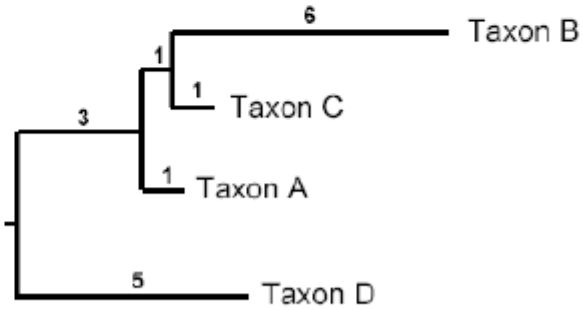
Genetic distance and evolutionary relationship

genetic distance: mutations accepted for site

evolutionary relationship: genetic connection during the time

similarity ≠ evolutionary relationship

B and C taxa are evolutionarily closer to each other (i.e., they have a most recent common ancestor) compared to the taxon A, though taxa C and A are more similar in sequence (the distance between B and C is equal to 7 = 6 + 1)

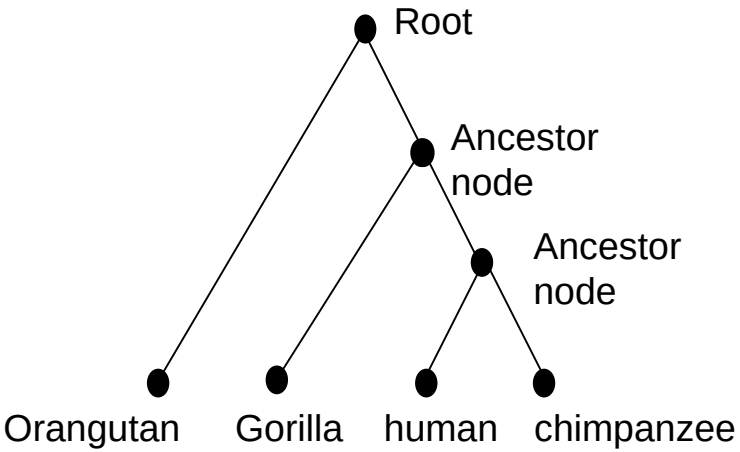


Phylogenetic Analysis

Problem of Phylogeny Reconstruction

Input: a set of species S (es. S = {Orangutan, Gorilla, …})

problem: to find the phylogeny T that represents the evolution of the species in S



Phylogenetic Analysis

Proteins or nucleic acids

In phylogeny they are used both:

Protein Sequences
- require 20x20 substitution matrix, very complex to deal with.
- are expression of coding regions.
- Identical amino acids can be expression of more codons

Nucleotide Sequences
- are described by 4x4 matrices.
- they can be extracted from genomic non-coding sequences, and then with a tendency to wider variation
- they have no degeneration or redundancy.

For the molecular phylogeny it is preferable to use nucleotide sequences

Phylogenetic Analysis

Assumptions

To calculate the evolutionary distance is necessary to formulate an evolutionary model!

it is therefore necessary to consider some general aspects that can be considered a-priori assumptions of the model:

1. all sites evolve independently
2. all sites change with the same chance
3. all replacements are equally likely
4. the replacement speed is constant over time
5. the composition of the bases is constant

the greater the number of a priori assumptions

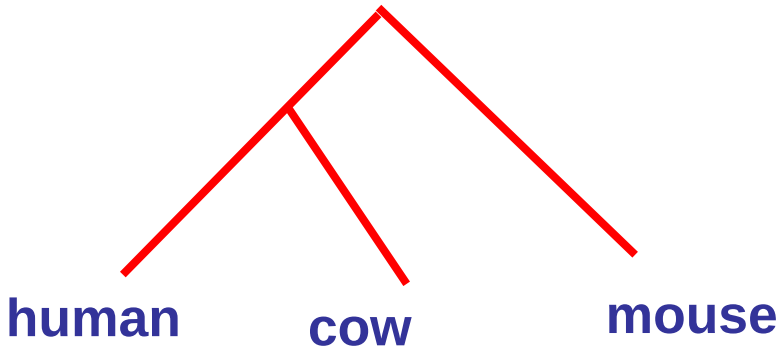
- the greater the simplicity of the model
- the lower the reliability of the results

Phylogenetic Analysis

Topology

It is defined as TOPOLOGY the general structure of a tree. If the branches do not give value to evolutionary distance, I have a cladogram, otherwise I have a FILOGRAM.

Trees WITH ROOT
accept as true the hypothesis of the molecular clock * and the nodes are in a precise temporal order.



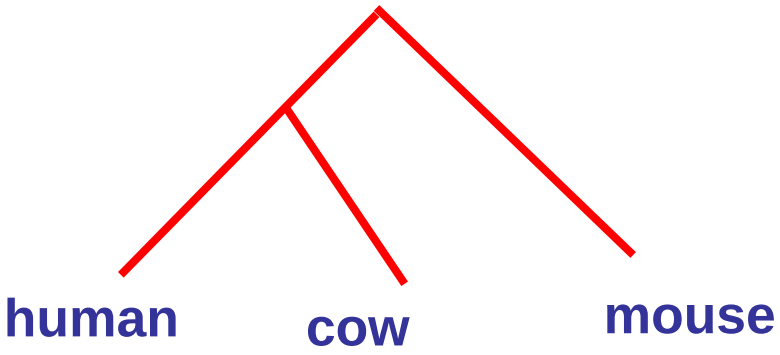
* The evolution is inevitably a divergent process and the number of mutations that accumulate over time is directly proportional to the time elapsed from the divergence of sequences in analysis. If this is true, given a genetic distance calculated by observing the differences, you can get the time from the moment when two sequences began to diverge.

Phylogenetic Analysis

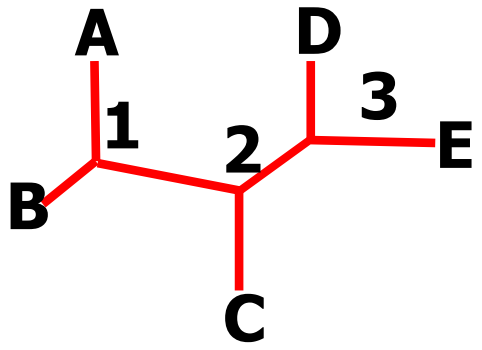
Topology

It is defined as TOPOLOGY the general structure of a tree. If the branches do not give value to evolutionary distance, I have a cladogram, otherwise I have a FILOGRAM.

Trees WITH ROOT
accept as true the hypothesis of themolecular clock * and the nodes are in a precise temporal order.



Trees WITHOUT ROOT
do not expect evolutionary significance in terms of time and simply describe the relationships between the sequences



Phylogenetic Analysis

The total number of trees that can be constructed with N sequences (called OTU, namely Operational Taxonomic Units) is given by:

Rooted $N_R = (2N - 3)! / (2^{N-3})*(N-3)!$

UnRooted $N_U = (2N - 5)! / (2^{N-3})*(N-3)!$

Phylogenetic Analysis

Methods for creating the trees

The algorithms to build trees can be distinguished according to the followed methodology:

- **Clustering algorithms** (Unweighted Pair Group Method with Arithmetic mean (UPMGA), Neighbor - Joining (NJ)) that are based on the observation of genetic distances calculated using multiple alignments.
- **Optimization algorithms** (Minimal evolution) that are algorithms for the Optimization of the trees on the basis of objective quality criteria.

Or according to origin of the data:

- **pre-calculated genetic distances**: are the most efficient since they have lower computing time.
- **Multi-aligned homologous Sequences :**that require much higher computation times.

Phylogenetic Analisys

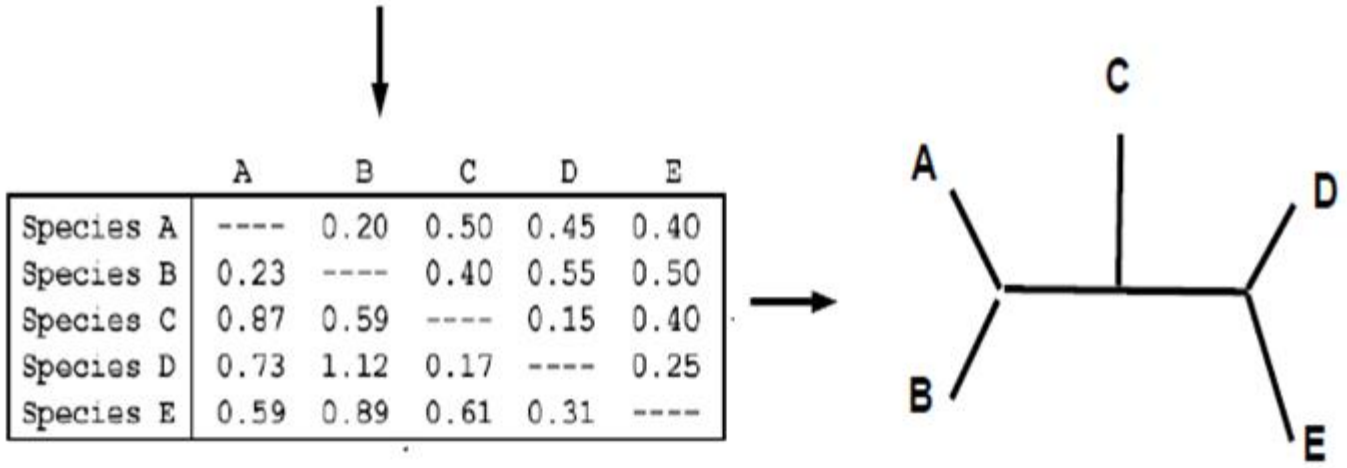
MEGA:
http://www.megasoftware.net

http://www.megasoftware.net/previousVersions.php

http://www.megasoftware.net/webhelp/helpfile.htm

Phylogenetic reconstruction systems based on distances

Taxa	Characters
Species A	ATGGCTATTCTTATAGTACG
Species B	ATCGCTAGTCTTATATTACA
Species C	TTCACTAGACCTGTGGTCCA
Species D	TTGACCAGACCTGTGGTCCG
Species E	TTGACCAGTTCTCTAGTTCG



	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	0.23	----	0.40	0.55	0.50
Species C	0.87	0.59	----	0.15	0.40
Species D	0.73	1.12	0.17	----	0.25
Species E	0.59	0.89	0.61	0.31	----

Estimation of genetic distances between sequences

genetic distances: the sequence data are transformed into matrices of distances using an evolutionary method

distances are calculated on the basis of **observed differences**

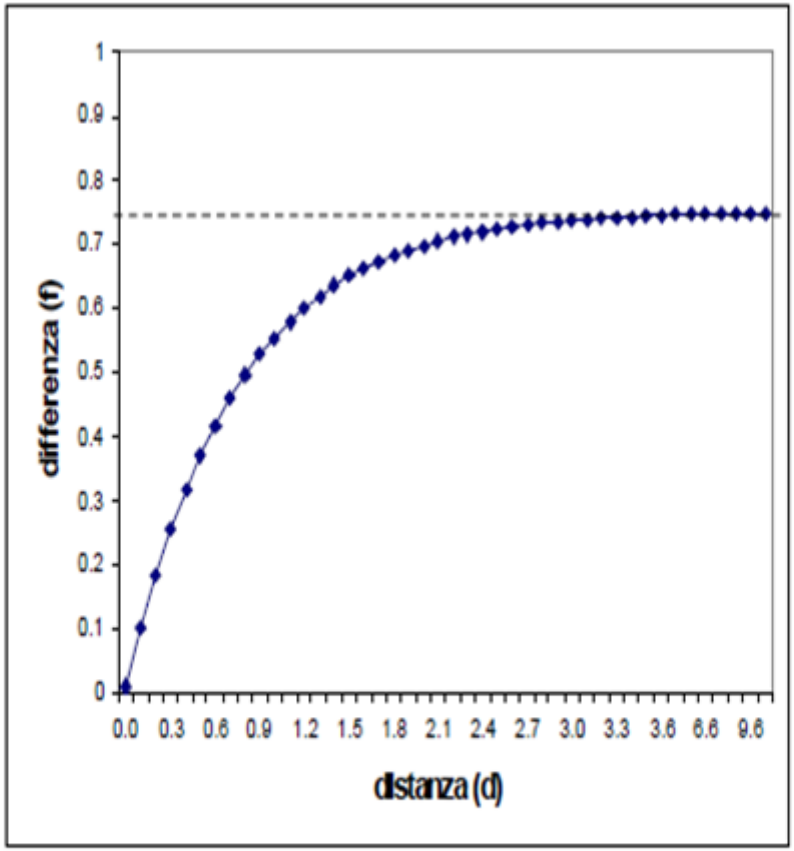
the calculation must take into account that **not all mutations are observable**

	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	0.23	----	0.40	0.55	0.50
Species C	0.87	0.59	----	0.15	0.40
Species D	0.73	1.12	0.17	----	0.25
Species E	0.59	0.89	0.61	0.31	----

incorrect Distance matrix (observed differences)

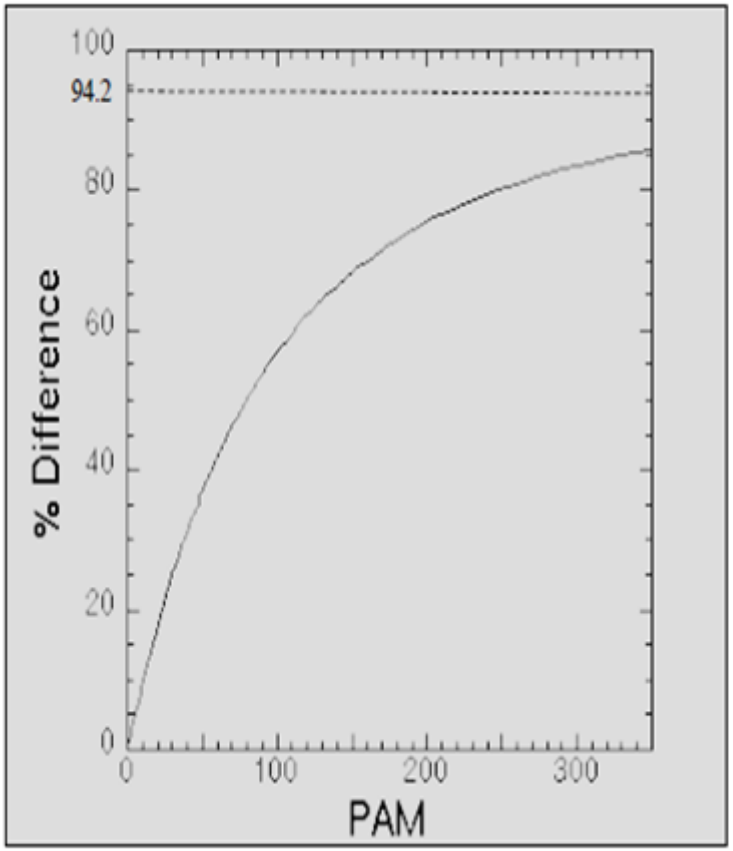
Correlation (estimate of the true number of mutations)

Computation of distances for nucleotide sequences



$d = -3/4 \ln(1 - 4/3\, f)$ Formula di Jukes-Cantor

Computation of distances for protein sequences



%Difference	PAM
1	1
5	5
10	11
15	17
20	23
25	30
30	38
35	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

PAM: mutations accepted in 100 sites

Clustering methods for genetic distances computation

Methods based on the distances

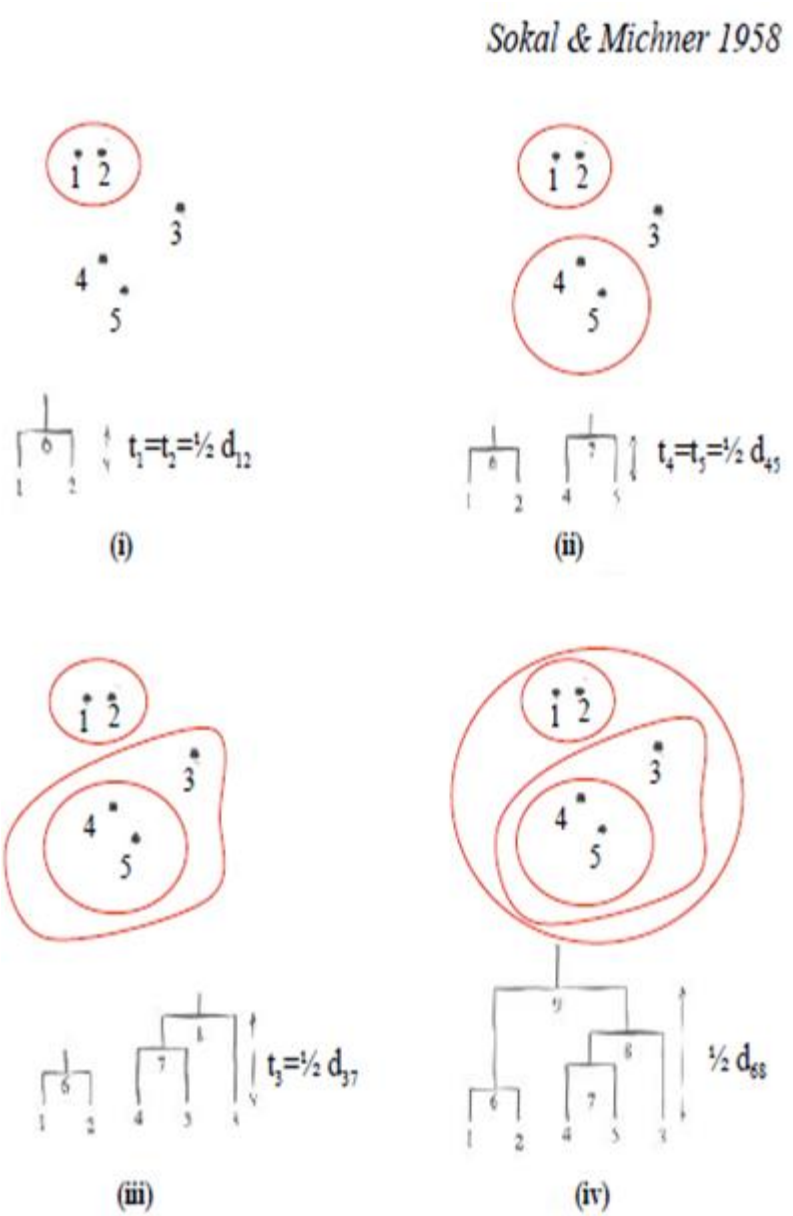
They use computation of distances + clustering method

- Neighbor-Joining
- UPGMA

UPGMA algorithm

UPGMA is a **clustering method** based on "Unweighted Pair Group Method using Arithmetic Average". It subsequently **groups** the sequences **from the most similar ones and gradually adding a node to the tree.** The **distances** between two taxa, between a node and a taxon, or between two nodes (i.e., the lengths of the edges) are given by the **arithmetic average of the distances.** The **tree can be imagined to be built from the bottom** upwards with each node added over the next. The last added node is the root.

UPGMA produces **rooted and ultra metric** trees. It can give trees with correct topology **only if the sequences comply with the molecular clock**



Neighbor Joining Algorithm

The Neighbor Joining method is a method for re-constructing phylogenetic trees, and computing the lengths of the branches of this tree.

In each stage, the two nearest nodes of the tree are chosen and defined as neighbors in our tree. This is done recursively until all of the nodes are paired together.

Neighbors are defined as a pair of OTU's (OTU=operational taxonomic units, or in other words – **leaves of the tree**), **who have one node connecting them**.

For instance, in the tree in figure 1, nodes A and B are neighbors (connected by only one internal node), and nodes C and D are neighbors, whereas nodes A and C (for example) are not neighbors.

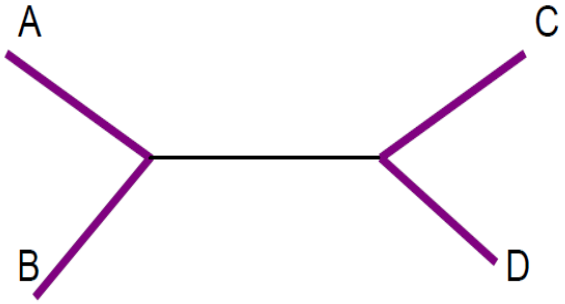
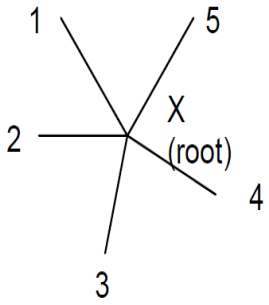


Figure 1

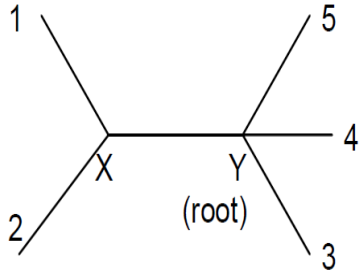
Neighbor Joining Algorithm
How do we find neighbors, and how de we construct our tree?

1. We start with a **star tree**:



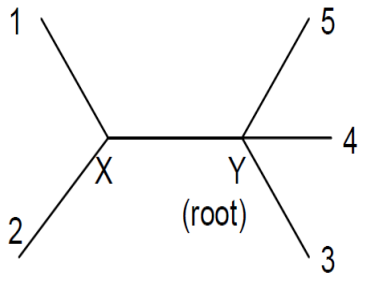
2. We define some kind of **distance** parameter between our nodes (1 through 5), and enter this parameter into a **distance matrix**. (The columns and rows of the matrix represent nodes, and the value i,j of the matrix represent the distance between node i and node j. Note that the matrix is symmetric, and that the diagonal is irrelevant, therefore only the top half (or lower half) are enough.)

3. **We pick the two nodes with the lowest value in the matrix** defined in step 2. These are **defined as neighbors.** For example, assuming nodes 1 and 2 are the nearest, we define them as neighbors

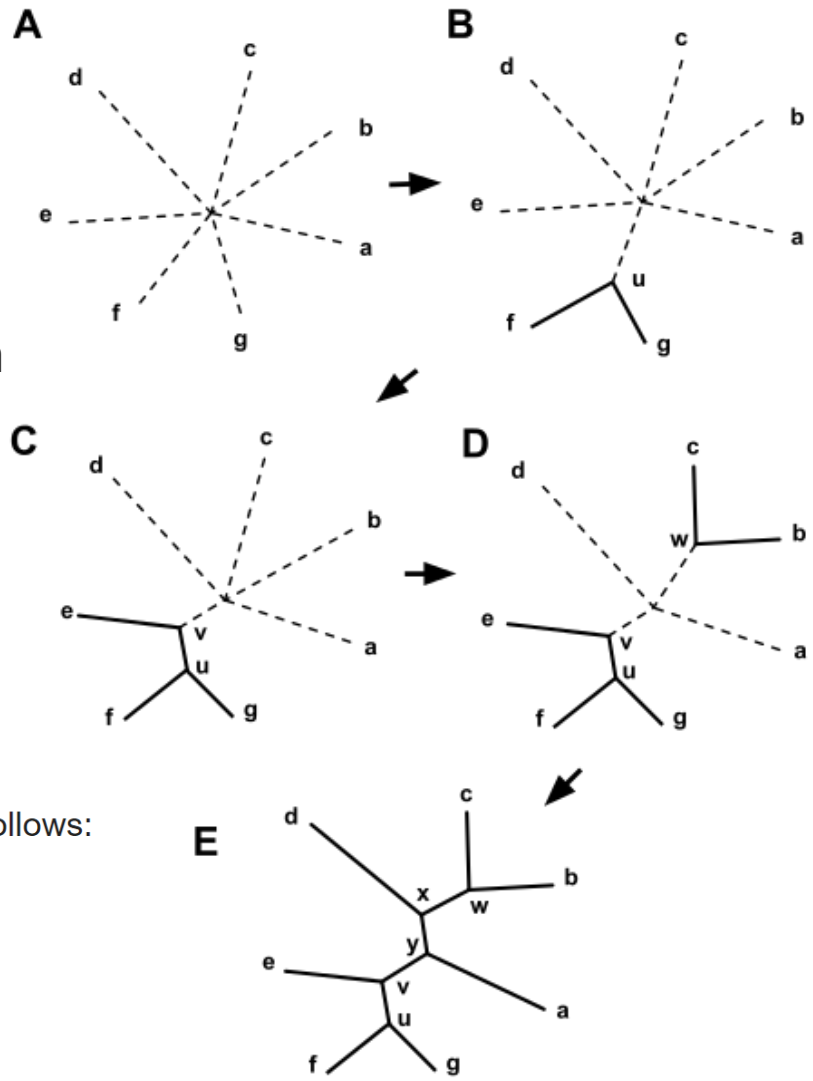


Neighbor Joining Algorithm

4. The new node we have added is defined as node X.
5. We now define the distance between node X and the rest of the nodes, and enter these distances into our distance matrix. We remove nodes 1 and 2 from our distance matrix.
6. We compute the branch lengths for the branches that have been joined (for figure 2(b), these are branches 1-X and 2-X) .
7. We repeat the process from stage 2 – once again we look for the 2 nearest nodes, and so on.



Starting with a star tree (A), the Q matrix is calculated and used to choose a pair of nodes for joining, in this case f and g. These are joined to a newly created node, u, as shown in (B). The part of the tree shown as solid lines is now fixed and will not be changed in subsequent joining steps.

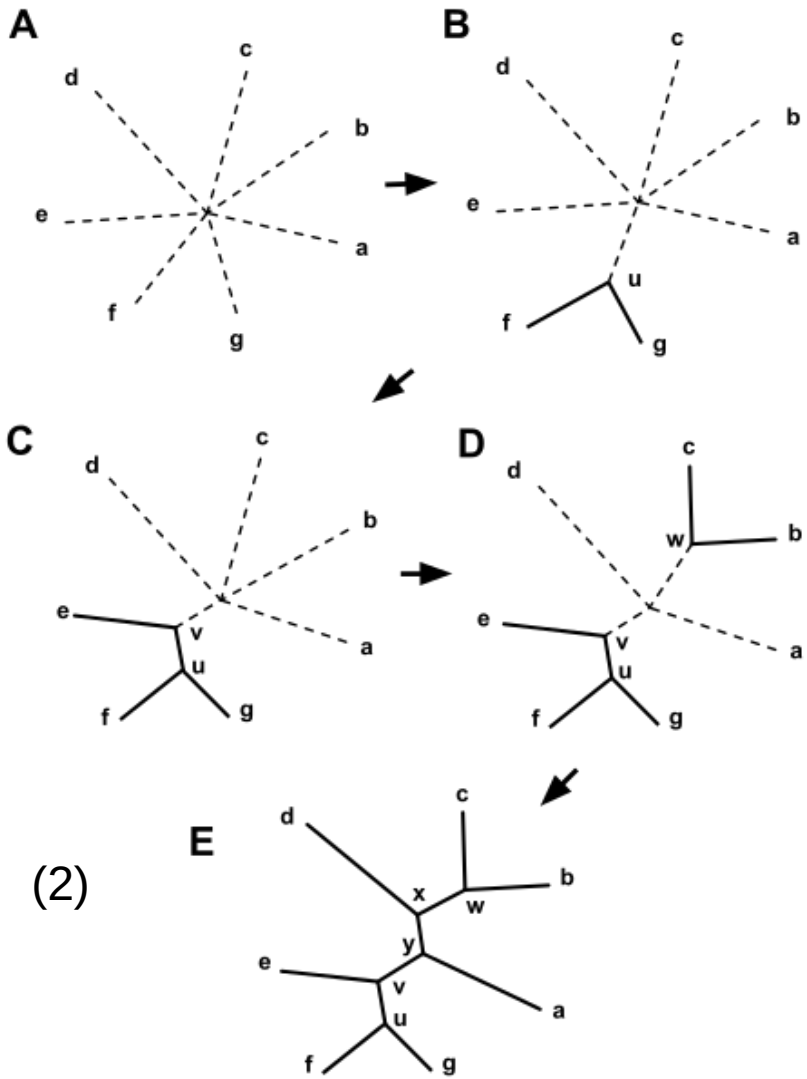


Based on a distance matrix relating the n taxa, calculate Q as follows:

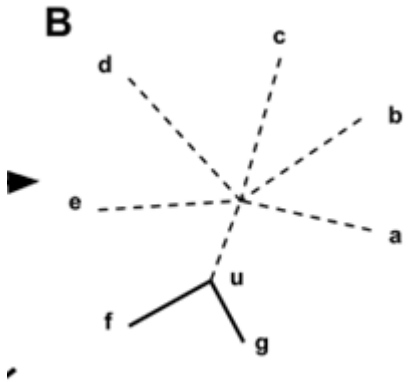
$$Q(i,j) = (n-2)d(i,j) - \sum_{k=1}^{n} d(i,k) - \sum_{k=1}^{n} d(j,k)$$

where $d(i,j)$ is the distance between taxa i and j.

The distances from node u to the nodes a-e are computed from equation (**1**). This process is then repeated, using a matrix of just the distances between the nodes, a,b,c,d,e, and u, and a Q matrix derived from it. In this case u and e are joined to the newly created v, as shown in (C). Two more iterations lead first to (D), and then to (E), at which point the algorithm is done, as the tree is fully resolved.



$$d(u,k) = \frac{1}{2}[d(f,k) + d(g,k) - d(f,g)] \quad (2)$$



Distance from the pair members to the new node

For each of the taxa in the pair being joined, use the following formula to calculate the distance to the new node:

$$\delta(f,u) = \frac{1}{2}d(f,g) + \frac{1}{2(n-2)}\left[\sum_{k=1}^{n} d(f,k) - \sum_{k=1}^{n} d(g,k)\right] \quad (2)$$

and:

$$\delta(g,u) = d(f,g) - \delta(f,u)$$

Taxa f and g are the paired taxa and u is the newly created node. The branches joining f and u and g and u, and their lengths, $\delta(f,u)$ and $\delta(g,u)$ are part of the tree which is gradually being created; they neither affect nor are affected by later neighbor-joining steps.

Optimization algorithms

Optimization algorithms (Minimal evolution) that are algorithms for the Optimization of the trees on the basis of objective quality criteria.

- Massima parsimonia
- Maximun Likelihood

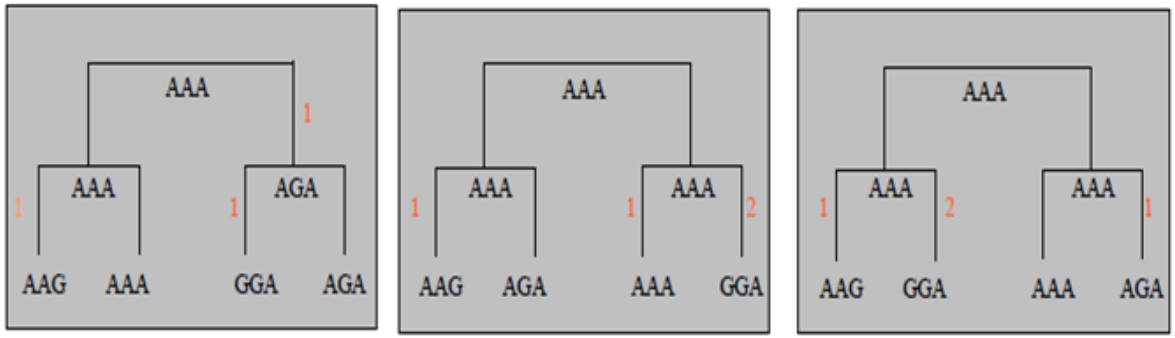
Massima Parsimonia

This method finds the unrooted tree that explains the considered sequences with the minimum number of substitutions. The method has two components:

1. Evaluate the cost of the tree in terms of mutations

2. Search among all possible tree the tree with lowest cost.

Seq1 AAG
Seq2 AAA
Seq3 GGA
Seq4 AGA



Considering the example, the algorithm will returns the tree at the left side since it has the lower cost (i.e., 3) with respect to the others that require 4 mutations

The maximum likelihood Tree

The method of maximum likelihood (ML) assesses the probability that the sequences were produced from a given evolutionary model and a given phylogenetic path. The tree (unrooted) that has the highest probability is called tree of ML and it is considered the optimal tree.

The algorithm has two components:

1. assessment of the likelihood that the relationships represented by a given tree (likelihood) is derived from each column of the alignment
2. Searches for all possible trees to find the tree with greater similarity

Advantages: solid probabilistic base
Disadvantages: it requires a high processing time, especially in the case of many taxa.

The maximum likelihood Tree

The maximum likelihood method uses standard statistical techniques for inferring probability distributions to assign probabilities to particular possible phylogenetic trees.

The method requires a substitution model to assess the probability of particular mutations; roughly, a tree that requires more mutations at interior nodes to explain the observed phylogeny will be assessed as having a lower probability.

This method is similar to the maximum-parsimony method, but maximum likelihood allows additional statistical flexibility by permitting varying rates of evolution across both lineages and sites.

In fact, the method requires that evolution at different sites and along different lineages must be statistically independent.

Procedure for the phylogenetic analysis of sequences

1. Selection of (homologous) sequences
2. Multiple alignment of the selected sequences
3. Execute the chosen method for the phylogenetic reconstruction
4. Visualization of the derived tree
5. Phylogenetic reconstruction validation

Validation of the phylogenetic prediction

There are two ways to estimate the degree of confidence of a specific phylogenetic reconstruction. And it is recommended to use both:

1. **Comparison of topologies obtained with different methods for the construction of the tree,** possibly one based on the distance and the other on the characters.

2. **Statistical estimation of the reliability of the results** obtained through random sampling the considering data (bootstrap)