



# Molecular Evolution and Phylogenetics ...a very short course


**Hernán J. Dopazo***
Pharmacogenomics and Comparative Genomics Unit

Bioinformatics Department†

Centro de Investigación Príncipe Felipe‡
**CIPF**

Valencia - Spain


2005


*hdopazo@ochoa.fib.es
†http://bioinfo.ochoa.fib.es
‡http:www.ochoa.fib.es/principal.htm

# 1. Introduction

## 1.1. Three basic questions

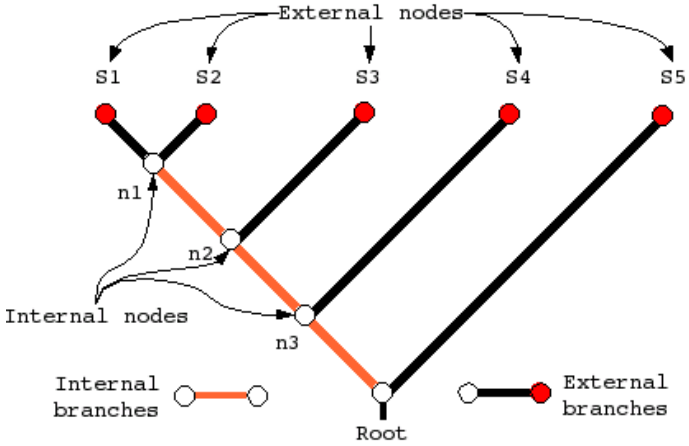
- Why use phylogenies?
  - Like astronomy, biology is an **historical** science!
  - The knowledge of the past is important to solve many questions related to biological patterns and processes.
- Can we know the past?
  - We can postulate alternative evolutionary scenarios **(hypothesis)**
  - Obtain the proper dataset and get statistical confidence
- What means to know **"...the phylogeny"**?
  - The ancestral-descendant relationships **(tree topology)**
  - The distances between them **(tree branch lengths)**

**Phylogenies are working hypotheses!!!**

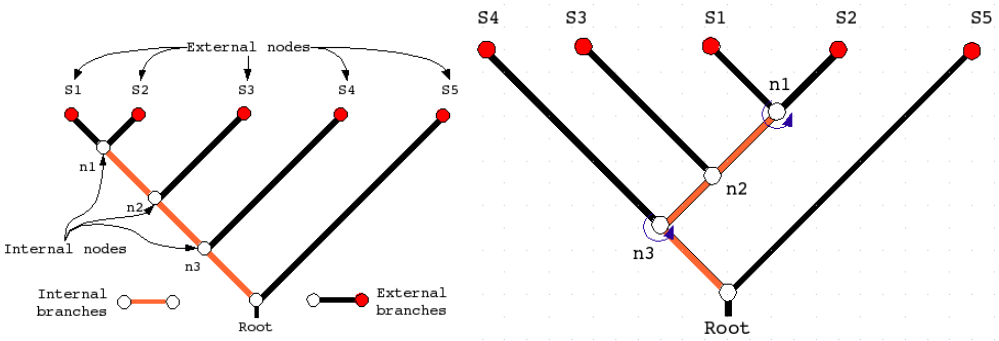
# 2. Tree Terminology

## 2.1. Topology, branches, nodes & root

- **Nodes & branches.** Trees contain internal and external nodes and branches. In molecular phylogenetics, **external nodes** are sequences representing **genes, populations or species!**. Sometimes, **internal nodes** contain the ancestral information of the clustered species. A **branch** defines the relationship between sequences in terms of descent and ancestry.

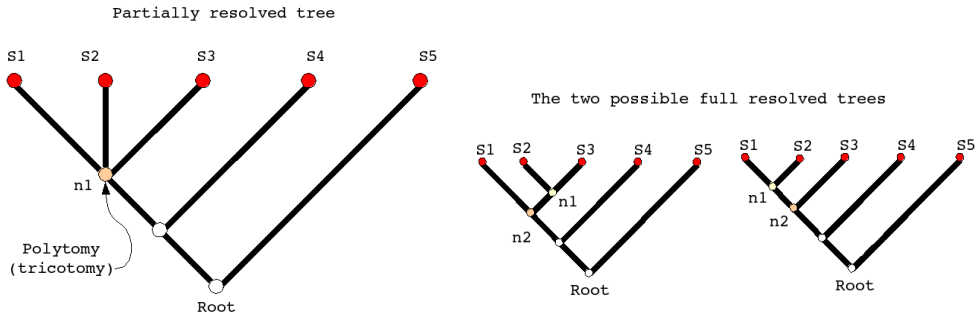



- **Root** is the common ancestor of all the sequences.
- **Topology** represents the branching pattern. Branches **can rotate** on internal nodes. Instead of the singular aspect, the folowing trees represent a single phylogeny.




**The topology is the same!!**

- **Taxa**. (*plural of taxon or operaqtional taxonomic unit (OTU)*) Any group of organisms, populations or sequences considered to be sufficiently distinct from other of such groups to be treated as a separate unit.
- **Polytomies**. Sometimes trees does not show fully bifurcated (binary) topologies. In that cases, the tree is considered **not resolved**. Only the relationships of species 1-3, 4 and 5 are known.

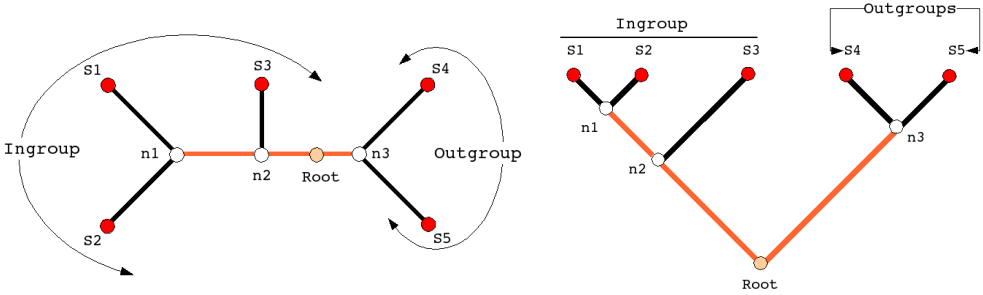



**Polytomies can be solved by using more sequences, more characters or both!!!**

## 2.2. Rooted & Unrooted trees

Trees can be **rooted** or **unrooted** depending on the explicit definition or not of **outgroup** sequence or taxa.

- **Outgroup** is any group of sequences used in the analysis that is not included in the sequences under study (**ingroup**).

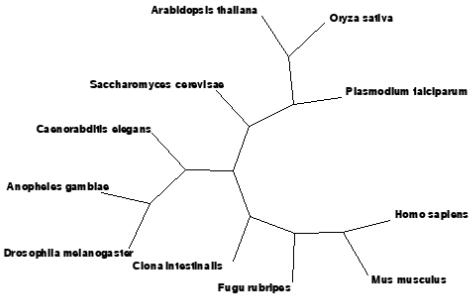



- **Unrooted trees** show the topological relationships among sequences althoug it is impossible to deduce wether nodes ($n_i$) represent a primitive or derived evolutionary condition.
- **Rooted trees** show the evolutionary basal and derived evolutionary relationships among sequences.

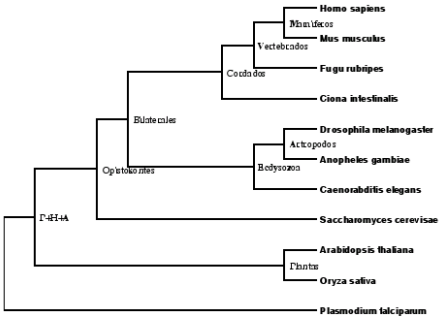
**Rooting by outgroup is frequent in molecular phylogenetics!!**

## 2.3. Cladograms & Phylograms

Trees showing branching order exclusivelly (**cladogenesis**) are principally the interest of systematists[1] to make inferences on taxonomy[2]. Those interesting in the evolutionary processes emphasize on branch lengths information (**anagenesis**).




Unrooted dendogram showing branching order

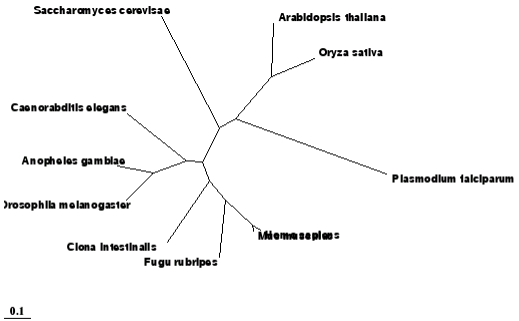



Rooted cladogram (cladistic methods)

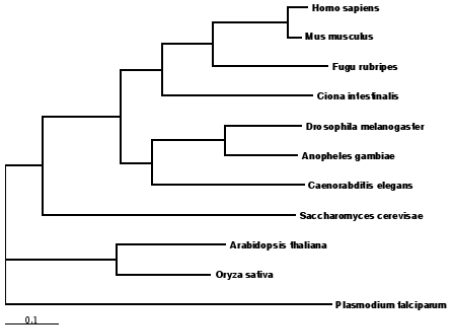
- **Dendrogram** is a branching diagram in the form of a tree used to depict degrees of relationship or resemblance.
- **Cladogram** is a branching diagram depicting the hierarchical arrangement of taxa defined by cladistic methods (the distribution of shared derived characters -synapomorphies-).

[1] The study of biological diversity.
[2] The theory and practice of describing, naming and classifying organisms

- **Phylogram** is a phylogenetic tree that indicates the relationships between the taxa and also conveys a sense of time or rate of evolution. The temporal aspect of a phylogram is missing from a cladogram or a generalized dendogram.
- **Distance scale** represents the number of differences between sequences (e.g. 0.1 means 10 % differences between two sequences)




Unrooted phylogram showing branch lengths



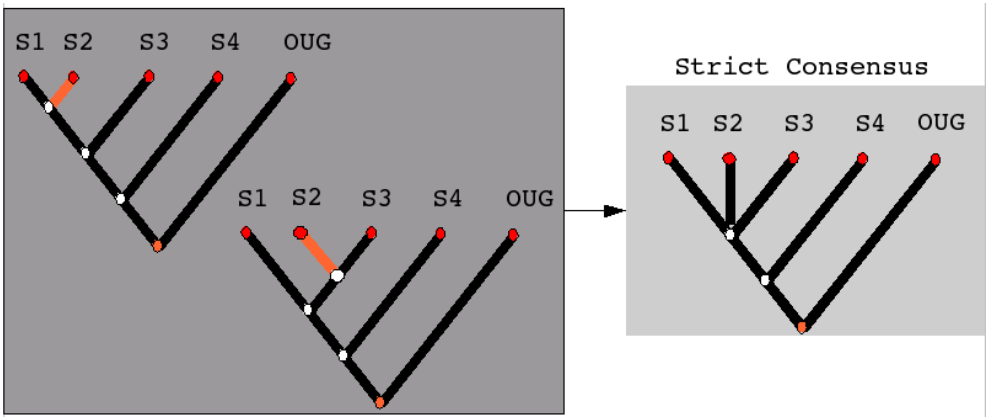

Unrooted phylogram

**Rooted and unrooted phylograms or cladograms are frequently used in molecular systematics!**

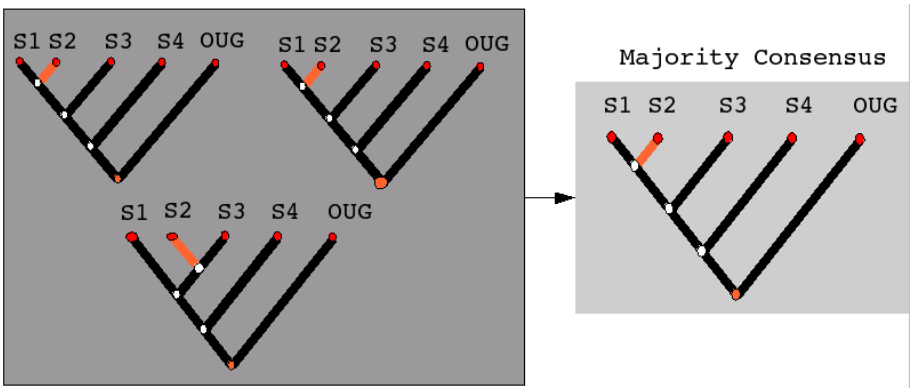
## 2.4. Consensus trees

It is frequent to obtain alternative phylogenetic hypothesis from a single data set. In such a case, it is usefull to summarize common or average relationships among the original set of trees. A number of different types of consensus trees have been proposed;

- The **strict consensus** tree includes only those monophyletic branches occurring in all the original trees. It is the most conservative consensus.



- The **majority rule consensus** tree uses a simple majority of relationships among the fundamental trees.




A consensus tree is a summary of how well the original trees agrees.

**A consensus tree is NOT a phylogeny!!.**[3]

A helpfull manual covering these and other concepts of the section can be obtained in [20, 12].

[3] Any consensus tree may be used as a phylogeny only if it is identical in topology to one of the original equally parsimonious trees.

# 3. Homology

**The Origin of Species. Charles Darwin. Chapter 14**

What can be more curious than that the hand of a man, formed for grasping, that of a mole for digging, the leg of the horse, the paddle of the porpoise, and the wing of the bat, should all be constructed on the same pattern, and should include similar bones, in the same relative positions?

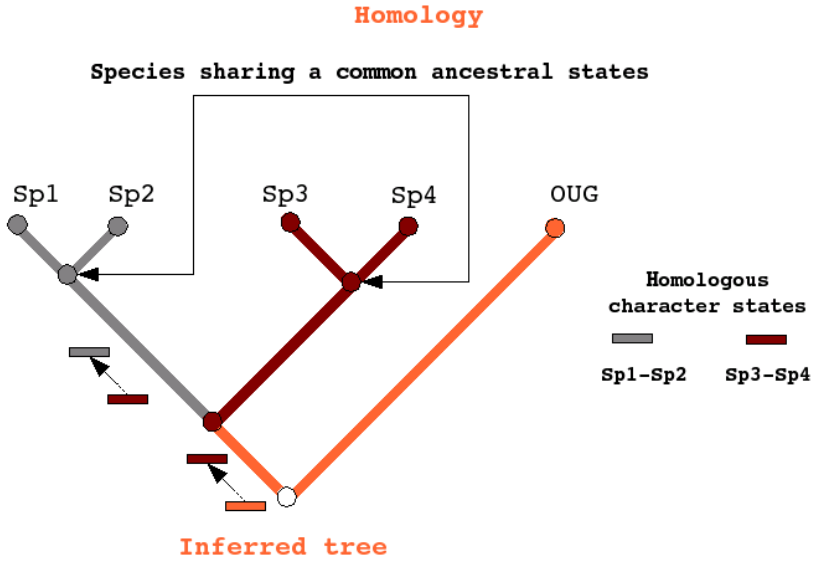
How inexplicable are the cases of serial homologies on the ordinary view of creation!

Why should similar bones have been created to form the wing and the leg of a bat, used as they are for such totally different purposes, namely flying and walking?

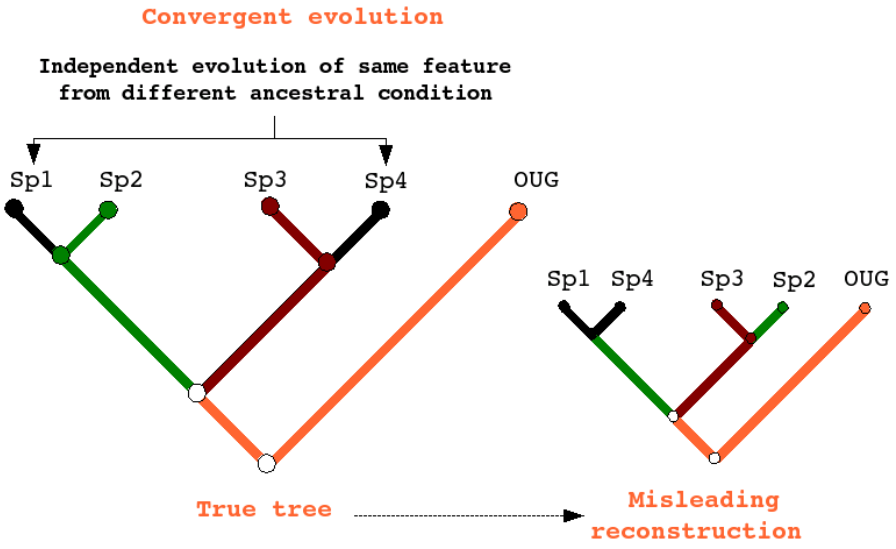
**Since Darwin homology was the result of descent with modification from a common ancestor.**

## 3.1. Homoplasy

- Similarity among species could represent true homology (just by sharing the same ancestral state) or, **homoplastic** events like **convergence, parallelism** or **reversals**;
- **Homology** is *a posteriori* tree construction definition.

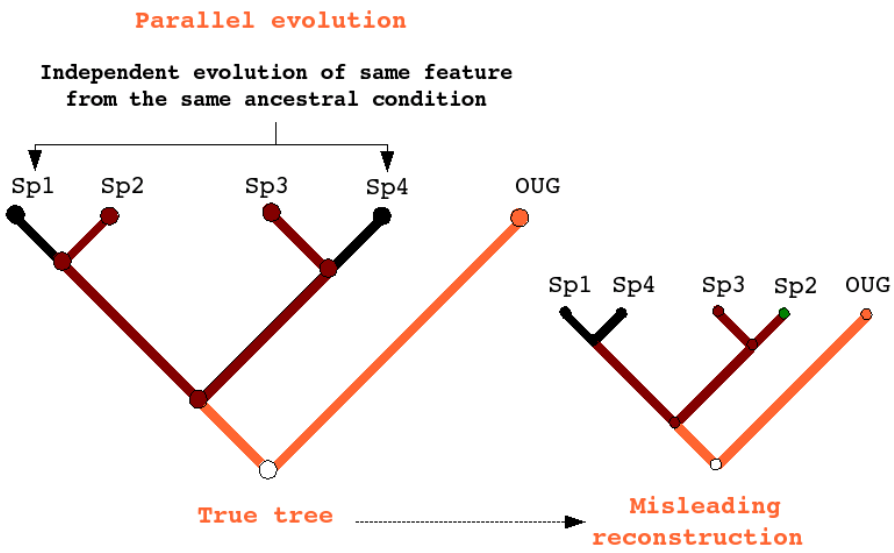



- **Convergences** are ...



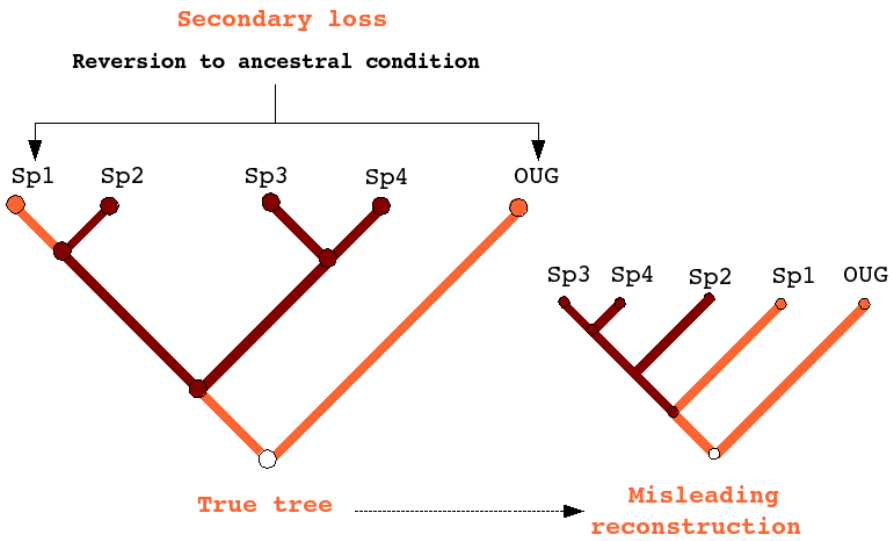

**Homoplasy** can provide misleading evidence of phylogenetic relationships!! (if mistakenly interpreted as homology).

- **Parallels** are ...




**Homoplasy** can provide misleading evidence of phylogenetic relationships!! (if mistakenly interpreted as homology).

- **Reversions** are ...




**Homoplasy** can provide misleading evidence of phylogenetic relationships!! (if mistakenly interpreted as homology).

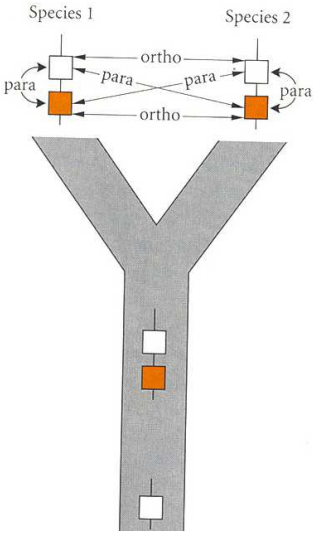
## 3.2. Similarity

- For molecular sequence data, **homology** means that two sequences or even two characters within sequences are descended from a common ancestor.
- This term is frequently mis-used as a synonym of **similarity**.
- as in **two sequences were 70% homologous**.
- **This is totally incorrect!**
- Sequences show a certain amount of similarity.
- From this similarity value, we can probably infer that the sequences are homologous or not.
- Homology is like pregnancy. You are either pregnant or not.
- Two sequences are either homologous or they are not.

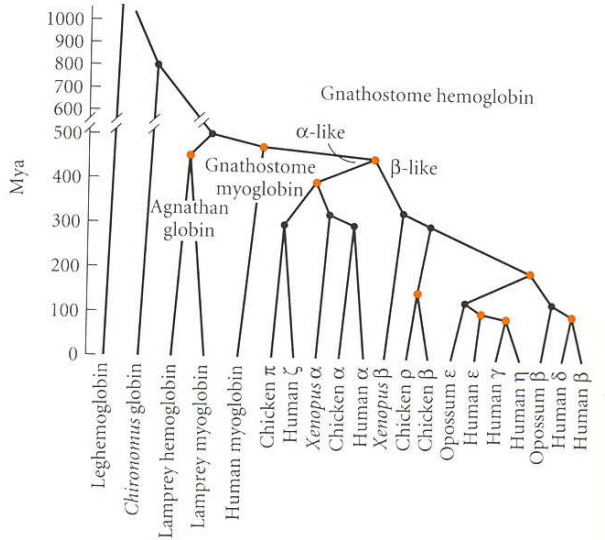
## 3.3. Sequence homology

In molecular studies it is important to distinguish among kinds of **homology**[6];

- **Ortholog**: Homologous genes that have diverged from each other after speciation events (e.g., human $\beta$- and chimp $\beta$-globin).
- **Paralog**: Homologous genes that have diverged from each other after gene duplication events (e.g., $\beta$- and $\gamma$-globin)




- **Xenolog**: Homologous genes that have diverged from each other after lateral gene transfer events (e.g., antibiotic resistance genes in bacteria).
- **Homolog**: Genes that are descended from a common ancestor (e.g., all globins).



- **Positional homology**: Common ancestry of specific amino acid or nucleotide positions in different genes.

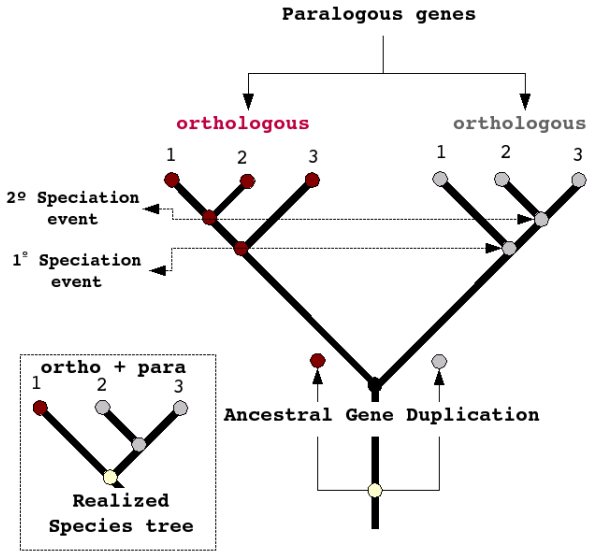
```
    11 50462
Homo.sapie VGLLGRTGSGKSTLLSAFLRLLNTEG-EIQI
Mus.muscul VGLLGRTGSGKSTLLSAFLRMLNIKG-DIEI
Fugu.rubri MGLLGRTGSGKSTLLSALLRLASTDG-EISI
Ciona.inte VGIVGRTGAGKSSLISTLFRLNEYSKGSVMI
Droso.mela VGIVGRTGAGKSSLIGALFRLAHIEG-EIFI
Anoph.gamb VGIVGRTGAGKSSLIGALFRLAQVEG-EIRL
Caeno.eleg VGIVGRTGAGKSSLTLALFRIIEADGGSIEI
Sacch.cere IGIVGRTGAGKSTIITALFRFLEPETGHIKI
Arabi.thal IGIVGRTGSGKTTLISALFRLVEPVGGKIVV
Oryza.sati IGVVGRTGSGKSTLVQALFRLVEPVEGHIIV
Plasm.falc IGIVGKSGAGKSTMILSILGLIGTTRGRITI
```

# 4. Molecular Evolution

## 4.1. Species & Genes trees

It is obvious that all phylogenetic reconstruction of sequences are **genes trees**. The naive expectation of molecular systematics is that phylogenies for genes match those of the organisms or species (**species trees**). *There are many reasons why this needs not be so!!.*

1. If there were **duplications**, (gene family) only the phylogenetic reconstruction of **orthologous** sequences could guarantize the expected[4] or true **species tree**.




[4]The expected tree is the tree that can be constructed by using infinitely long sequences

## 4.2. Molecular clock

The **molecular clock hypothesis** postulates that for any given macromolecule (a protein or DNA sequence), the rate of evolution -*measured as the mean number of amino acids or nucleotide sequence change per site per year*- is approximately constant over time in all the evolutionary lineages [21].

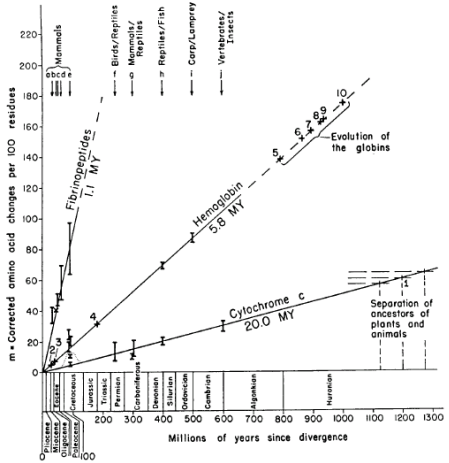


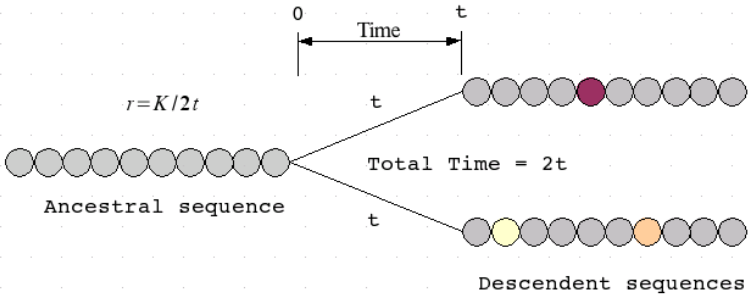

Fig. 8.3. Rates of amino acid substitution in the fibrinopeptides, hemoglobin, and cytochrome *c*. Comparisons for which no adequate time coordinate is available are indicated by numbered crosses. Point 1 represents a date of 1200 ± 75 MY (million years) for the separation of plants and animals, based on a linear extrapolation of the cytochrome *c* curve. Points 2–10 refer to events in the evolution of the globin family. The $\delta/\beta$ separation is at point 3, $\gamma/\beta$ is at 4, and $\alpha/\beta$ is at 500 MY (carp/lamprey). From Dickerson (1971).

This hypothesis has estimulated much interest in the use of macromolecules in evolutionay studies for two reasons:

- Sequences can be used as molecular markers to **date** evolutionary events.
- The degree of rate change among sequences and lineages can provide insights on **mechanisms** of molecular evolution. For example, a large increase in the rate of evolution in a protein in a particular lineage may indicate adaptive evolution.

## Substitution rate estimation

It is based on the number of aa substitution (distance) and divergence time (fossil calibration),




## There is no universal clock

It is known that **clock variation** exists for:

- different molecules, *depending on their functional constraints*,
- different regions in the same molecule,

Rates of amino acid substitution at the surface and heme pocket regions of the hemoglobin $\alpha$- and $\beta$-chains (Kimura and Ohta, 1973b).

| Region | $\alpha$-chain | $\beta$-chain |
|---|---|---|
| Surface | 1.4 (18) | 2.7 (23) |
| Heme pocket | 0.17 (19) | 0.24 (21) |

Note: The rate represents 'per amino acid site per year'. The values in the table should be multiplied by $10^{-9}$. The figures in brackets are the number of amino acid sites involved.

- different base position (synonimous-nonsynonimous),

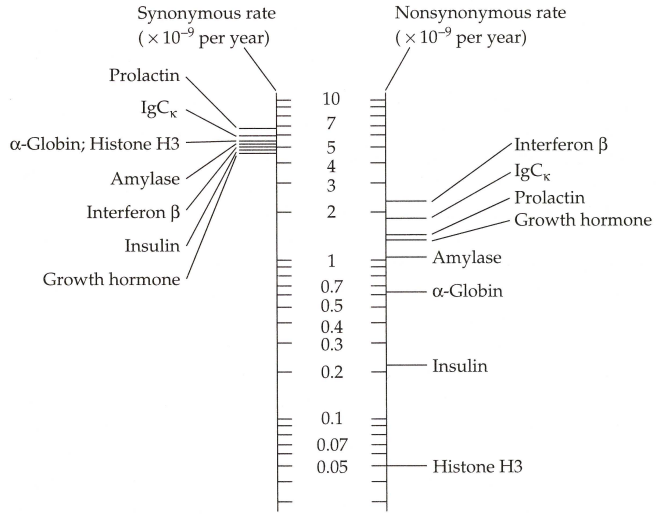



**Figure 8.14** Comparison of rates of synonymous and nonsynonymous nucleotide substitutions. Synonymous rates are generally much faster and much more uniform than nonsynonymous rates. (From Kimura 1986.)

- different genomes in the same cell,
- different regions of genomes,
- different taxonomic groups for the same gene (**lineage effects**)

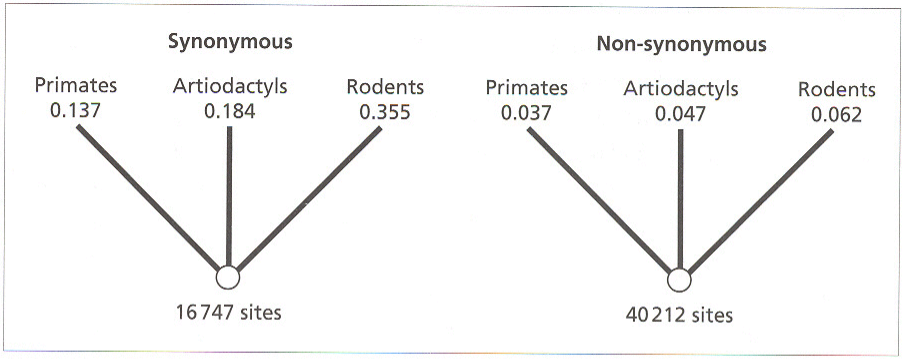
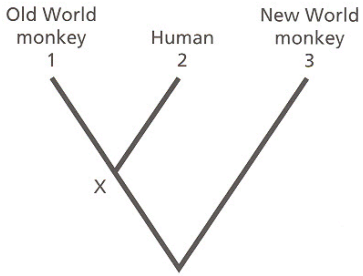



**Fig. 7.14** Numbers of synonymous and non-synonymous substitutions for 49 genes from three mammalian orders: primates, rodents and artiodactyls, the phylogenetic relationships of which approximate a 'star phylogeny'. Note that, in both cases, rodents have accumulated more substitutions than primates or artiodactyls. Adapted from Ohta (1995).

## Relative Rate Test

How to test the molecular clock?[5]




**Results**

(a) Synonymous sites in nine nuclear genes (3520 bp)
$d_{12} = 6.7$
$d_{13} - d_{23} = 2.3 \pm 0.6$*

(b) ψη–globin pseudogene (1827 bp)
$d_{12} = 7.9$
$d_{13} - d_{23} = 1.5 \pm 0.4$*

(c) Three introns (3376 bp)
$d_{12} = 6.9$
$d_{13} - d_{23} = 1.0 \pm 0.5$

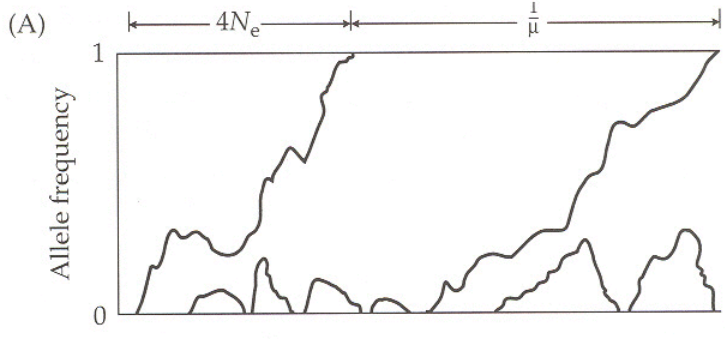
(d) Two flanking regions (936 bp)
$d_{12} = 7.9$
$d_{13} - d_{23} = 3.1 \pm 1.1$*

[5] See [13] and download RRtree!!

## 4.3. Neutral theory of evolution

At molecular level, the most frequent changes are those involving fixation in populations of neutral selective variants [8].

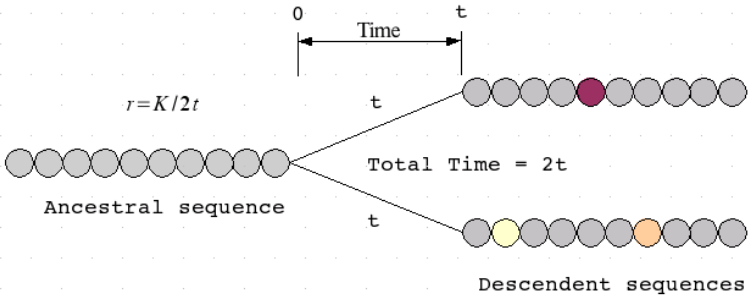
- Allelic variants are functionaly equivalent
- Neutralism does not deny adaptive evolution
- Fixation of new allelic variants occurs at a constant rate $\mu$.
- This rate does not depends on any other population parameter, then it's **like a clock!!** $2N\mu * 1/2N = \mu$



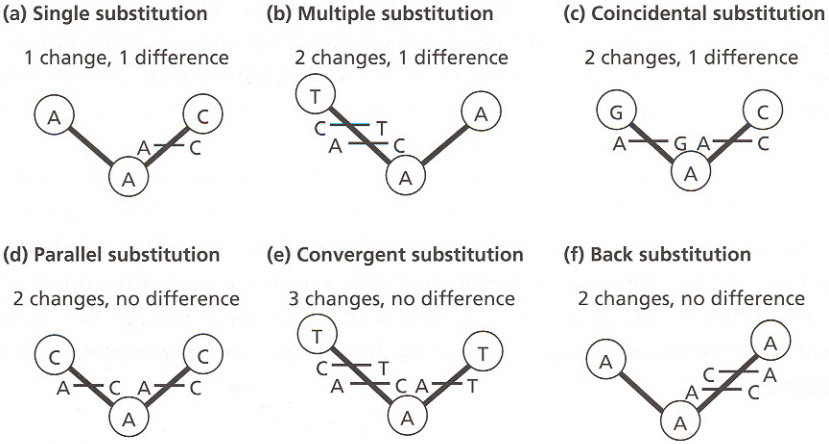

# 5. Evolutionary Models

## 5.1. Multiple Hits

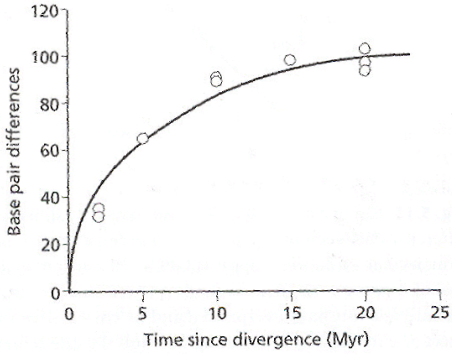
- The mutational change of DNA sequences varies with region. Even considering protein coding sequence alone, the patterns of nucleotide substitution at the first, second or third codon position are not the same.
- When two DNA sequences are derived from a common ancestral sequence, the descendant sequences gradually diverge by nucleotide substitution.
- A simple measure of sequence divergence is the proportion $p = N_d/N_t$ of nucleotide sites at which the two sequences are different.

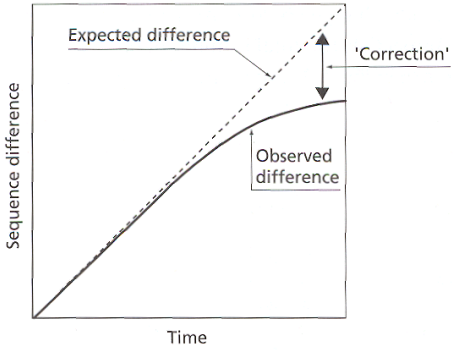



- When $p$ is large, it gives an underestimate of the number of of substitutions, because it does not take into account **multiple substitutions**.



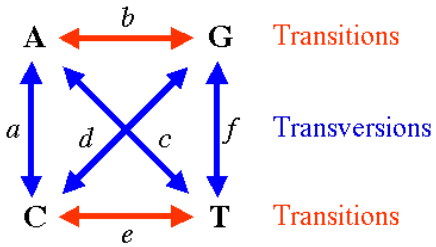

- Sequences may saturate due to multiple changes (**hits**) at the same position after lineage splitting.
- In the worst case, data may become random and all the **phylogenetic information** about relationships can be lost!!!



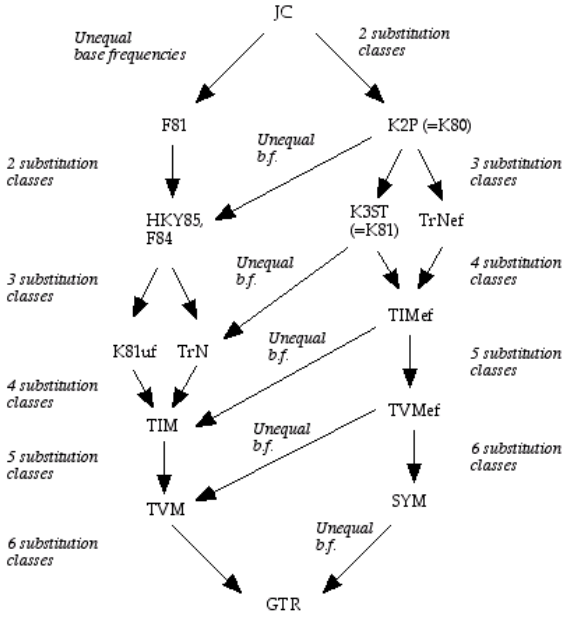

## 5.2. Models of nucleotide substitution

- In order to estimate **the number of nucleotide substitutions ocurred** it is necessary to use a mathematical model of nucleotide substitution. The model would consider the nucleotide frequencies and the instantaneous rate's change among them.




| Designation | Rate params | Base frequencies | Number of free params |
|---|---|---|---|
| JC | a=b=c=d=e=f | $\pi_A = \pi_C = \pi_G = \pi_T$ | 1 |
| K80, K2P | a=c=d=f, b=e | $\pi_A = \pi_C = \pi_G = \pi_T$ | 2 |
| TrNef | a=c=d=f, b, e | $\pi_A = \pi_C = \pi_G = \pi_T$ | 3 |
| K81, K3ST | a=f, b=e, c=d | $\pi_A = \pi_C = \pi_G = \pi_T$ | 3 |
| TVMef | a, c, d, f, b=e | $\pi_A = \pi_C = \pi_G = \pi_T$ | 5 |
| TIMef | a=f, c=d, b, e | $\pi_A = \pi_C = \pi_G = \pi_T$ | 4 |
| SYM | a, b, c, d, e, f | $\pi_A = \pi_C = \pi_G = \pi_T$ | 6 |
| F81 | a=b=c=d=e | $\pi_A, \pi_C, \pi_G, \pi_T$ | 4 |
| HKY | a=c=d=f, b=e | $\pi_A, \pi_C, \pi_G, \pi_T$ | 5 |
| TrN | a=c=d=f, b, e | $\pi_A, \pi_C, \pi_G, \pi_T$ | 6 |
| K81uf | a=f, b=e, c=d | $\pi_A, \pi_C, \pi_G, \pi_T$ | 6 |
| TVM | a, c, d, f, b=e | $\pi_A, \pi_C, \pi_G, \pi_T$ | 8 |
| TIM | a=f, c=d, b, e | $\pi_A, \pi_C, \pi_G, \pi_T$ | 7 |
| GTR, REV | a, b, c, d, e, f | $\pi_A, \pi_C, \pi_G, \pi_T$ | 9 |

- Interrrelationships among models for estimating the number of nucleotide substitutions among a pair of DNA sequences




- For constructing phylogenetic trees from distance measures, sophisticated distances are not neccesary more efficient.

Table 3.3 Observed numbers of the 10 pairs of nucleotides between the DNA sequences for the human and Rhesus monkey mitochondrial cytochrome *b* genes.

| | Transition | | Transversion | | | | Identical Pair | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Codon Position | TC | AG | TA | TG | CA | CG | TT | CC | AA | GG | $n_d$ | Total ($n$) |
| First | 21 | 22 | 5 | 1 | 5 | 4 | 68 | 93 | 100 | 56 | 58 | 375 |
| Second | 20 | 3 | 6 | 1 | 0 | 2 | 140 | 87 | 71 | 45 | 32 | 375 |
| Third | 60 | 16 | 6 | 5 | 49 | 2 | 11 | 122 | 102 | 2 | 138 | 375 |
| All | 101 | 41 | 17 | 7 | 54 | 8 | 219 | 302 | 273 | 103 | 228 | 1125 |

*Note:* The numbers at the first, second, and third codon positions are shown separately.

- Indeed, by using sophisticated models distances show higher variance values.

Table 3.4 Estimates ($\hat{d}$) of the number of nucleotide substitutions per site between the human and Rhesus monkey mitochondrial cytochrome *b* genes for the first, second, and third codon positions ($\hat{d} \times 100$).

| Position in Codon | p | Jukes-Cantor | Kimura | Tajima-Nei | Tamura-Nei |
|---|---|---|---|---|---|
| First | 15.5 ± 1.9 | 17.3 ± 2.4 | 17.8 ± 2.5 | 18.0 ± 2.6 | 17.9 ± 2.5 |
| Second | 8.5 ± 1.4 | 9.1 ± 1.6 | 9.2 ± 1.7 | 9.2 ± 1.7 | 9.3 ± 1.7 |
| Third | 36.8 ± 2.5 | 50.6 ± 4.9 | 52.3 ± 5.4 | 66.5 ± 9.4 | 87.9 ± 39.0 |

- Of course, corrected distances are greather than the observed.

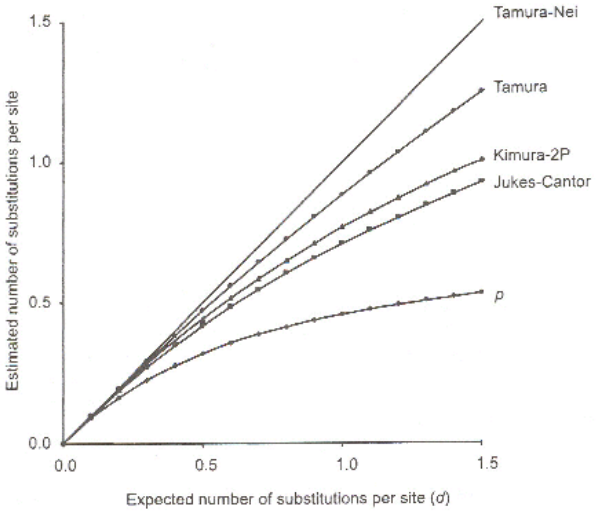



FIGURE 3.1. Estimates of the number of nucleotide substitutions obtained by different distance measures when actual nucleotide substitution follows the Tamura-Nei model. The nucleotide frequencies assumed are $g_A = 0.3$, $g_T = 0.4$, $g_C = 0.2$, and $g_G = 0.1$; and the two transition/transversion rate ratios assumed are $\alpha_1/\beta = 4$ and $\alpha_2/\beta = 8$.

**Distance correction methods share several assumptions:**

- All nucleotide sites change independently.
- The substitution rate is constant over time and in different lineages
- The base composition is at equilibrium (all sequences have the same base frequencies)
- The conditional probabilities of nucleotide substitutions are the same for all sites and do not change over time.

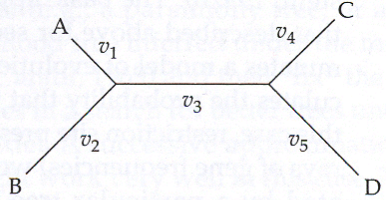
**While these assumptions make the methods tractable, they are in many cases unrealistic.**

## 6. Distance Methods

**Distance matrix methods is a major family of phylogenetic methods trying to fit a tree to a matrix of pairwise distance** [1, 5]. Distance are generally corrected distances.

- The best way of thinking about distance matrix methods is to consider distances as estimates of the branch length separating that pair of species.
- Branch lengths are not simply a function of time, they reflect expected amounts of evolution in different branches of the tree.
- Two branches may reflect the same elapsed time (sister taxa), but they can have different expected amounts of evolution.
- The product $r_i * t_i$ is the branch length
- The main distance-based tree-building methods are **cluster analysis, least square** and **minimum evolution**.
- They rely on different assumptions, and their success or failure in retrieving the correct phylogenetic tree depends on how well any particular data set meet such assumptions.

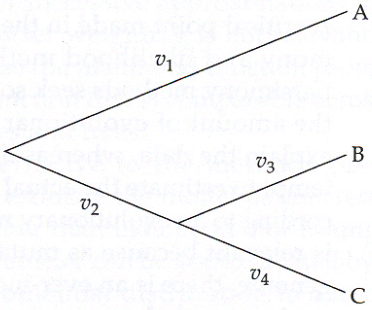
(A)




Additive properties:

$$d_{AB} = v_1 + v_2$$
$$d_{AC} = v_1 + v_3 + v_4$$
$$d_{AD} = v_1 + v_3 + v_5$$
$$d_{BC} = v_2 + v_3 + v_4$$
$$d_{BD} = v_2 + v_3 + v_5$$
$$d_{CD} = v_4 + v_5$$

(B)




Additive properties:

$$d_{AB} = v_1 + v_2 + v_3$$
$$d_{AC} = v_1 + v_2 + v_4$$
$$d_{BC} = v_3 + v_4$$

Ultrametric properties:

$$v_3 = v_4$$
$$v_1 = v_2 + v_3 = v_2 + v_4$$

## 6.1. Cluster Analysis

Cluster analysis derived from clustering algorithms popularized by Sokal and Sneath[16]

### 6.1.1. UPGMA

One of the most popular distance approach is the **unweighted pair-group method with arithmetic mean (UPGMA)**, which is also the simplest method for tree reconstruction [10].

1. Given a matrix of pairwise distances, find the clusters (taxa) $i$ and $j$ such that $d_{ij}$ is the minimum value in the table.
2. Define the depth of the branching between $i$ and $j$ ($l_{ij}$) to be $d_{ij}/2$
3. If $i$ and $j$ are the last 2 clusters, the tree is complete. Otherwise, create a new cluster called $u$.
4. Define the distance from $u$ to each other cluster ($k$, with $k \neq i$ or $j$) to be an average of the distances $d_{ki}$ and $d_{kj}$
5. Go back to step 1 with one less cluster; clusters $i$ and $j$ are eliminated, and cluster $u$ is added.

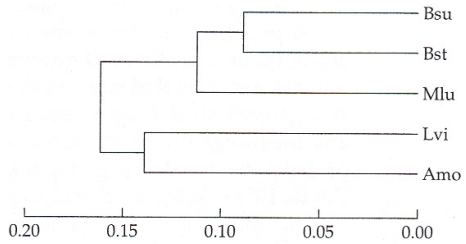
   The variants of UPGMA are in the step 4. Weighted PGMA(WPGM::$d_{ku} = d_{ki}+d_{kj}/2$). Complete linkage ($d_{ku} = max(d_{ki}, d_{kj})$. Single linkage($d_{ku} = min(d_{ki}, d_{kj})$.

| | Bsu | Bst | Lvi | Amo | Mlu |
|---|---|---|---|---|---|
| Bsu | — | **0.1715** | 0.2147 | 0.3091 | 0.2326 |
| Bst | | — | 0.2991 | 0.3399 | 0.2058 |
| Lvi | | | — | 0.2795 | 0.3943 |
| Amo | | | | — | 0.4289 |
| Mlu | | | | | — |

| | Bsu-Bst | Lvi | Amo | Mlu |
|---|---|---|---|---|
| Bsu-Bst | — | 0.2569 | 0.3245 | **0.2192** |
| Lvi | | — | 0.2795 | 0.3943 |
| Amo | | | — | 0.4289 |
| Mlu | | | | — |

| | Bsu-Bst-Mlu | Lvi | Amo |
|---|---|---|---|
| Bsu-Bst-Mlu | — | 0.3027 | 0.3593 |
| Lvi | | — | **0.2795** |
| Amo | | | — |

| | Bsu-Bst-Mlu | Lvi-Amo |
|---|---|---|
| Bsu-Bst-Mlu | — | **0.3310** |
| Lvi-Amo | | — |




The smallest distance in the first table is 0.1715 substitutions per sequence position separating *Bacillus subtilis* and *B. stearothermophilus*. The distance between Bsu-Bst to Lvi (*Lactobacillus viridescens*) is (0.2147+0.2991)/2=0.2569. In the second table, joins Bsu-Bst to Mlu(*Micrococcus luteus*) at the depth 0.1096(=0.2192/2). The distances Bsu-Bst-Mlu to Lvi is (2*0.2569+0.3943)/3=0.3027. Notice that this value is identical to (Bsu:Lvi+Bst:Lvi+Mlu:Lvi)/3. Each taxon in the original data table contributes equally to the averages, this is why the method called **unweighted**

**UPGMA method supposes a cloclike behaviour of all the lineages, giving a rooted and ultrametric tree.**

### 6.1.2. NJ (Neighboor Joining)

A variety of methods related to cluster analysis have been proposed that will correctly reconstruct additive trees, whether the data are ultrametric or not. NJ removes the assumption that the data are ultrametric.

1. For each terminal node $i$ calculate its net divergence $(r_i)$ from all the other taxa using $\mapsto r_i = \sum_{k=1}^{N} d_{ik}$ [6].
2. Create a rate-corrected distance matrix $(\mathbf{M})$ in which the elements are defined by $\mapsto M_{ij} = d_{ij} - (r_i + r_j)/(N-2)$ [7].
3. Define a new node $u$ whose three branches join nodes $i, j$ and the rest of tree. Define the lengths of the tree branches from $u$ to $i$ and $j \mapsto v_{iu} = d_{ij}/2 + ((r_i - r_j)/[2(N-2)]; v_{ju} = d_{ij} - v_{iu}$
4. Define the distance from $u$ to each other terminal node (for all $k \neq i$ or $j) \mapsto d_{ku} = (d_{ik} + d_{jk} - d_{ij})/2$
5. Remove distances to nodes $i$ and $j$ from the matrix, decrease $N$ by 1
6. If more than2 nodes remain, go back to step 1. Otherwise, the tree is fully defined except for the length of the branch joining the two remaining nodes ($i$ and $j$) $\mapsto v_{ij} = d_{ij}$

[6] N is the number of terminal nodes

[7] Only the values $i$ and $j$ for which $M_{ij}$ is minimum need to be recorded, saving the entire matrix is unnecessary

The main virtue of neighbor-joining is its efficiency. It can be used on very large data sets for which other phylogenetic analysis are computationally prohibitive.

| | Bsu | Bst | Lvi | Amo | Mlu | R | R/3 |
|---|---|---|---|---|---|---|---|
| Bsu | — | 0.1715 | 0.2147 | 0.3091 | 0.2326 | 0.9279 | 0.3093 |
| Bst | –0.4766 | — | 0.2991 | 0.3399 | 0.2058 | 1.0163 | 0.3388 |
| Lvi | –0.4905 | –0.4356 | — | **0.2795** | 0.3943 | 1.1876 | 0.3959 |
| Amo | –0.4527 | –0.4514 | **–0.5689** | — | 0.4289 | 1.3574 | 0.4525 |
| Mlu | –0.4972 | –0.5535 | –0.4221 | –0.4441 | — | 1.2616 | 0.4205 |

Lvi to node 1 distance = 0.2795/2 + (0.3959 – 0.4525)/2 = 0.1114
Amo to node 1 distance = 0.2795 – 0.1114 = 0.1681

| | Bsu | Bst | Mlu | Node 1 | R | R/2 |
|---|---|---|---|---|---|---|
| Bsu | — | 0.1715 | 0.2326 | **0.1222** | 0.5263 | 0.2631 |
| Bst | –0.3701 | — | 0.2058 | 0.1798 | 0.5571 | 0.2785 |
| Mlu | –0.3856 | –0.4278 | — | 0.2719 | 0.7103 | 0.3551 |
| Node 1 | **–0.4278** | –0.3856 | –0.3701 | — | 0.5739 | 0.2869 |

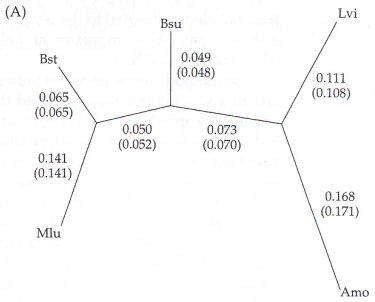
Bsu to node 2 distance = 0.1222/2 + (0.2631 – 0.2869)/2 = 0.0492
node 1 to node 2 distance = 0.1222 – 0.0492 = 0.0730

| | Bst | Mlu | Node 2 | R | R/1 |
|---|---|---|---|---|---|
| Bst | — | 0.2058 | **0.1146** | 0.3204 | 0.3204 |
| Mlu | –0.5116 | — | 0.1912 | 0.3970 | 0.3970 |
| Node 2 | **–0.5116** | –0.5116 | — | 0.3058 | 0.3058 |

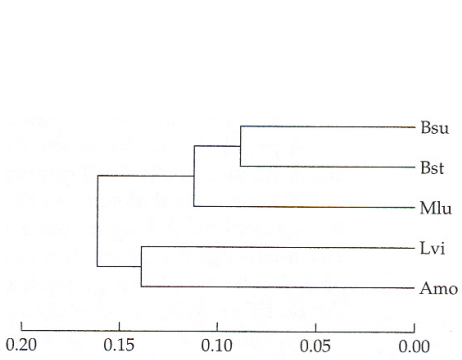
Bst to node 3 distance = 0.1146/2 + (0.3204 – 0.3058)/2 = 0.0646
node 2 to node 3 distance = 0.1146 – 0.0646 = 0.0500

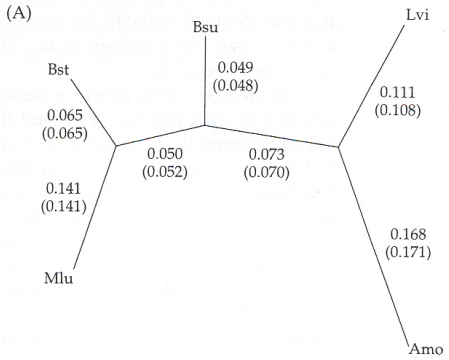
| | Mlu | Node 3 |
|---|---|---|
| Mlu | — | 0.1412 |
| Node 3 | | — |

Mlu to node 3 distance = 0.1412




**Unlike the UPGMA, NJ does not assume that all lineages evolve at the same rate and produces an unrooted tree.**




## 6.2. Pros & Cons of Distance Methods

- **Pros**:
  - They are very fast,
  - There are a lot of models to correct for multiple,
  - LRT may be used to search for the best model.
- **Cons**:
  - Information about evolution of particular characters is lost

# 7. Maximum Parsimony

Most biologists are familiar with the usual notion of **parsimony** in science, which essentially maintains that simpler hypotheses are prefereable to more complicated ones and that *ad hoc* hypotheses should be avoided whenever possible. The principle of *maximum parsimony* (MP) searches for a tree that requires **the smallest number of evolutionary changes** to explain differences observed among OTUs.

In general, parsimony methods operate by selecting trees that minimize the total tree length: **the number of evolutionary steps (transformation of one character state to another) require to explain a given set of data**.

In mathematical terms: from the set of possible trees, find all trees $\tau$ such that $L_{(\tau)}$ is **minimal**

$$L_{(\tau)} = \sum_{k=1}^{B} \sum_{j=1}^{N} w_j . diff(x_{k'j}, x_{k''j})$$

Where $L_{(\tau)}$ is the length of the tree, $B$ is the number of branches, $N$ is the number of characters, $k'$ and $k''$ are the two nodes incident to each branch $k$, $x_{k'j}$ and $x_{k''j}$ represent either element of the input data matrix or optimal character-state assignments made to internal nodes, and diff$(y, z)$ is a function specifying the cost of a transformation from state $y$ to state $z$ along any branch. The coefficient $w_j$ assigns a weight to each character. Note also that diff$(y, z)$ needs not to be equal diff$(z, y)$.[8]

[8] For methods that yield unrooted trees diff$(y, z)$ =diff$(z, y)$.

Determining the length of the tree is computed by algorithmic methods[4, 15]. However, we will show how to calculate the length of a particular tree topology ((W,Y),(X,Z))[9] for a specific site of a sequence, using Fitch (A) and transversion parsimony (B)[10]:

Seq. W ....ACA**G**GAT...
Seq. X ....ACA**C**GCT...
Seq .Y ....GTA**A**GGT...
Seq. Z ....GCA**C**GAC...

(A)

$$\text{equal} = \begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

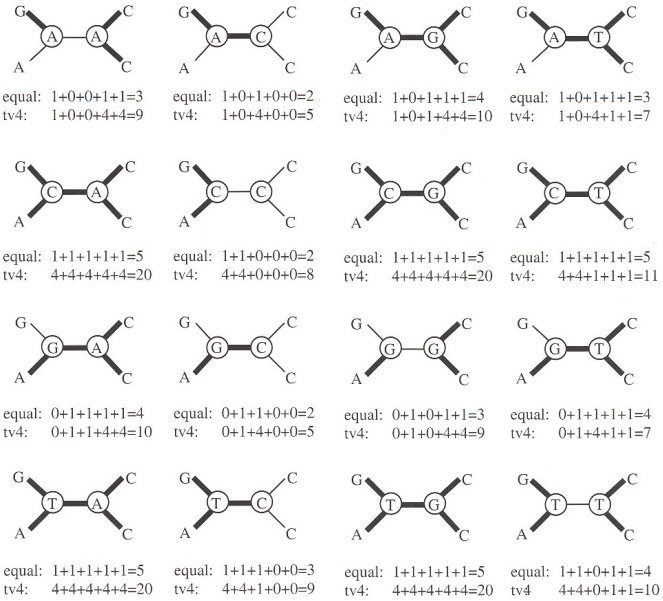
(B)

$$\text{tv4} = \begin{bmatrix} 0 & 4 & 1 & 4 \\ 4 & 0 & 4 & 1 \\ 1 & 4 & 0 & 4 \\ 4 & 1 & 4 & 0 \end{bmatrix}$$

- With equal costs, the minimum is 2 steps, achieved by 3 ways (internal nodes "A-C", "C-C", "G-C"),
- The alternative trees ((W,X),(Y,Z)) and ((W,Z),(Y,X)) also have 2 steps,
- Therefore, the character is said to be **parsimony-uninformative**,[11]
- With 4:1 ts:tv weighting scheme, the minimum length is 5 steps, achived by two reconstructions (internal nodes "A-C" and"G-C"),
- By evaluating the alternative topologies finds a minimum of 8 steps,

[9] Newick format
[10] Matrix character states: A,C,G,T
[11] A site is informative, only it favors one tree over the others

- Therefore, under unequal costs, the character **becomes informative**. The use of unequal costs may provide more information for phylogenetic reconstruction,



## 7.1. Pros & Cons of MP

- **Pros**:
  - Does not depend on an explicit model of evolution,
  - At least gives both, a tree and the associated hypotheses of character evolution,
  - If homoplasy is rare, gives reliable results,
- **Cons**:
  - May give misleading results if homplasy is common (*Long branch attraction effect*)
  - Underestimate branch lengths
  - Parsimony is often justified by phylosophical, instead statistical grounds.

# 8. Searching Trees

## 8.1. How many trees are there?

The obvious method for searching the most parsimonious tree is to consider all posible trees, one after another, and evaluate them. We will see that this procedure becomes impossible for more than a few number of taxa (∼11). Felsenstein [2] deduced that:

$$B(T) = \prod_{i=3}^{T} (2i - 5)$$

An unrooted, fully resolved tree has:

- $T$ terminal nodes, $T - 2$ internal nodes,
- $2T - 3$ branches; $T - 3$ interior and $T$ peripheral,
- $B(T)$ alternative topologies,
- Adding a **root**, adds one more **internal node** and one more **internal branch**,
- Since the root can be placed along any $2T - 3$ branches, the number of possible **rooted trees** becomes,

$$B(T) = (2T - 3) \prod_{i=3}^{T} (2i - 5)$$

| OTUs | Rooted trees | Unrooted trees |
|---|---|---|
| **2** | 1 | 1 |
| **3** | 3 | 1 |
| **4** | 15 | 3 |
| **5** | 105 | 15 |
| **6** | 954 | 105 |
| **7** | 10,395 | 954 |
| **8** | 135,135 | 10,395 |
| **9** | 2,027,025 | 135,135 |
| **10** | 34,459,425 | 2,027,025 |
| **11** | $> 654\text{x}10^{6}$ | $> 34\text{x}10^{6}$ |
| **15** | $> 213\text{x}10^{12}$ | $> 7\text{x}10^{12}$ |
| **20** | $> 8\text{x}10^{21}$ | $> 2\text{x}10^{20}$ |
| **50** | $> 6\text{x}10^{81}$ | $> 2\text{x}10^{76}$ |

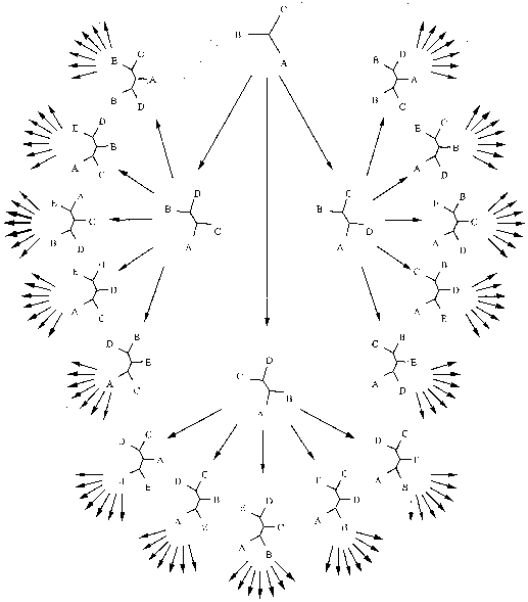
**The observable universe has about $8.8\text{x}10^{77}$ atoms**

**There is not memory neither time to evaluate all the trees!!**

For 11 or fewer taxa, a brute-force **exhaustive search** is feasible!!
For more than 11 taxa an **heuristic search** is the best solution!!

## 8.2. Exhaustive search methods

- Every possible tree is examined; **the shortest tree will always be found**,
- Taxon addition sequence is important only in that **the algorithm needs to remember where it is**,
- Search will also generate **a list** of the lenths of all possible trees, which can be plotted as an histogram,



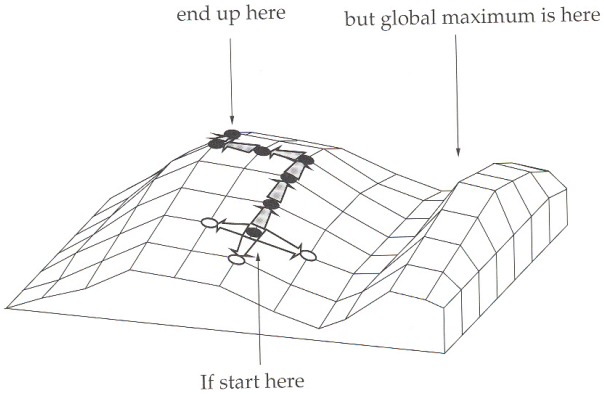
## 8.3. Heuristic search methods

When a data set is **too large to permit the use of exact methods**, optimal trees must be sought via heuristic approaches that **sacrifice the guarantee of optimality in favor of reduced computing time**

Two kind of algorithms can be used:

1. Greedy Algorithms
2. Branch Swapping Algorithms

### 8.3.1. Greedy Algorithms




Strategies of this sort are often called *the greedy algorithm* because they seize the first improvement that they see. Two major algorithms exist:

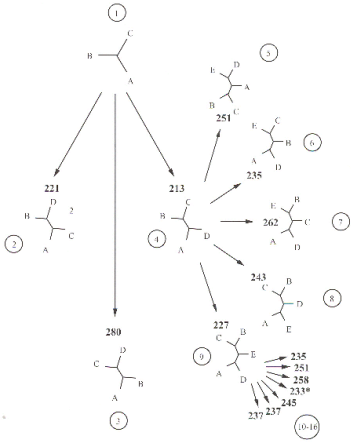
- Stepwise Addition,
- Star Decomposition[12]

**Both algoritms are prone to entrapment in local optima**

[12]The most common star decomposition method is the NJ algorithm

## Stepwise Addition

- Use addition sequence similar to that for an exhaustive search, but at each addition, determines the shortest tree, and add the next taxon to that tree.
- Addition sequence will affect the tree topology that is found!




A greedy stepwise-addition search applied to the example in Figure 7.2. The best four-taxon tree is determined by evaluating the lengths of the three trees obtained by joining Taxon D to Tree 1 containing only the first three taxa. Taxa E and F are then connected to the five and seven possible locations, respectively, on Trees 4 and 9, with only the shortest trees found during each step being used for the next step. In this example, the 233-step tree obtained is not a global optimum (see Figure 7.2). Circled numbers indicate the order in which phylogenetic trees are evaluated in the stepwise-addition search.

## Star Decomposition

- Start with all taxa in an unresolved (star) tree,
- Form pairs of taxa, and determine length of tree with paired taxa.

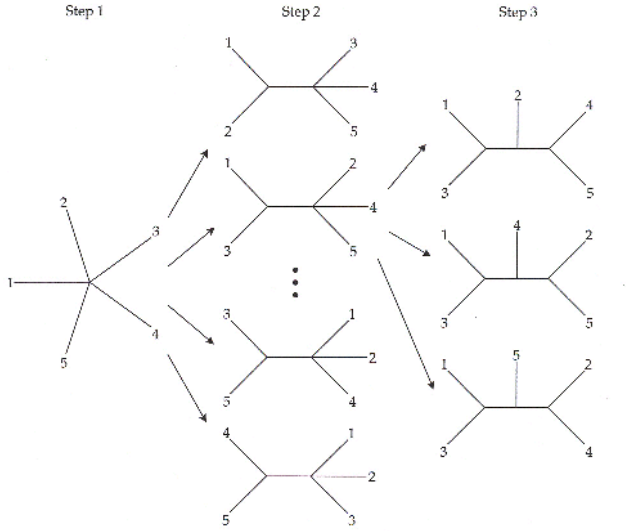



**Figure 25** Heuristic tree selection using star decomposition method. At each step, the optimality criterion is evaluated for each possible joining of a pair of lineages leading away from the central node. The best tree found during each step becomes the starting point for the next step.

### 8.3.2. Branch Swapping Algorithms

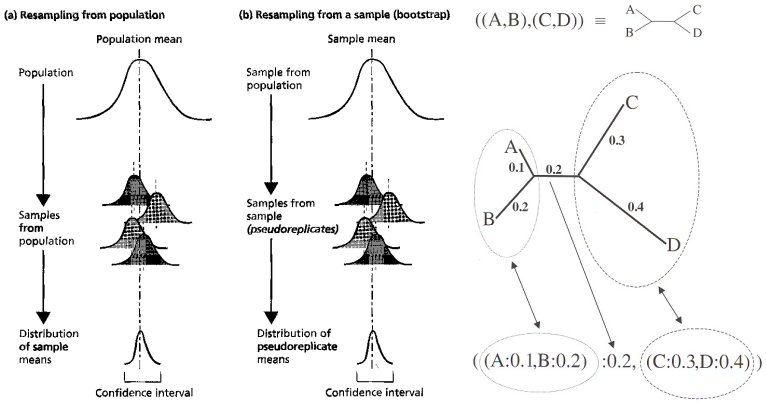
It may be possible to improve the *greedy* solutions by performing sets of predefined rearrangements, or branch swappings. Examples of branch swapping algorithms are:

- NNI - Nearest Neighbor Interchange,
- SPR - Subtree Pruning and Regrafting,
- TBR - Tree Bisection and Reconnection.

# 9. Tree Confidence

## 9.1. Non-parametric bootstrapping

- For many simple distributions there are simple equations for calculating confidence intervals around an estimate (e.g., std error of the mean)
- Trees, however are rather complicated structures, and it is extremely difficult to develop equations for confidence intervals around a phylogeny.
- One way to measure the confidence on a phylogenetic tree is by means of the **bootstrap** non-parametric method of resampling the same sample many times.



- Each sample from the original sample is a **pseudoreplicate**. By generation many hundred or thousand pseudoreplicates, a *majority consensus rule tree* can be obtained.
- High bootstrap values $> 90\%$ is indicative of strong **phylogenetic signal**.
- Bootstrap can be viewed as a way of exploring the robustness of phylogenetic inferences to perturbations
- **Jackkniffe** is another non-parametric resampling method that differentiates from bootstrap in the way of sampling. Some proportion of the characters are randomly selected and deleted (withouth replacement).
- Another technique used exclusively for parsimony is by means of **Decay index** or **Bremmer support**. This is the length difference between the shortest tree including the group and the shortest tree excluding the group (The extra-steps required to overturn a group.[13]
- **DI** & **BPs** generally correlates!!

[13] See [19] for a practical example using PAUP*[17]

# 10. PC Lab

## 10.1. Download Programs

- PHYLIP 3.6 http://evolution.genetics.washington.edu/phylip.html
- MEGA 3.0 http://www.megasoftware.net
- TREE-PUZZLE http://www.tree-puzzle.de/
- MODELTEST http://darwin.uvigo.es/
- MrBayes http://morphbank.ebc.uu.se/mrbayes/download.php
- TreeView http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

# 11. Phylogenetic Links

- Software:
  - **The Felsenstein node** http://evolution.genetics.washington.edu/phylip/software.html
  - **The R. Page Lab.** http://taxonomy.zoology.gla.ac.uk/software/software.html
- Courses:
  - **Molecular Systematics and Evolution of Microorganisms.** http://www.dbbm.fiocruz.br/james/index.html
  - **Workshop on Molecular Evolution** http://workshop.molecularevolution.org/
  - **P. Lewis MCB/EEB Course** http://www.eeb.uconn.edu/Courses/EEB372/
- Tools:
  - **Clustalw at EBI** http://www.ebi.ac.uk/clustalw/
  - **Phylip Web** http://cbrmain.cbr.nrc.ca:8080/cbr/jsp/ServicePage_e.jsp?id=38
  - **Phylip Doc**http://www.hgmp.mrc.ac.uk/Registered/Help/phylip/phylip.html

# 12. Credits

This presentation is based on:[14]

- Major Book or Chapters References:
  - Swofford, D. L. *et al.* **1996**. Phylogenetic inference [18].
  - Harvey, P. H. *et al.* **1996**. New Uses for New Phylogenies [7].
  - Li, W. S. **1997** . Molecular Evolution [9].
  - Page, R. & Holmes, E. **1998**. Molecular evolution. A phylogenetic approach [7].
  - Nei, M. & Kumar, S. **1999** . Molecular evolution and phylogenetics [11].
  - Salemi, M. & Vandamme, A. (ed.) **2003**. The phylogenetic handbook [14].
  - Felsenstein, J. **2004**. Inferring phylogenies [3] .
- On Line Phylogenetic Resources:
  - http://www.dbbm.fiocruz.br/james/index.html .**Molecular Systematics and Evolution of Microorganisms**. The Natural History Museum, London and Instituto Oswaldo Cruz, FIOCRUZ.
  - Peter Foster's "The Idiot's Guide to the Zen of Likelihood in a Nutshell in Seven Days for Dummies" at http://www.bioinf.org/molsys/data/idiots.pdf
- Slides Production:
  - Latex and pdfscreen package.

[14] HJD take responsibility for innacuracies of this presentation.

## References


[1] L. L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetic Analysis: Models and estimation procedures. *American Journal of Human Genetics*, 19:223–257, 1967.

[2] J. Felsenstein. The number of evolutionary trees. (Correction:, Vol.30, p.122, 1981). *Syst. Zool.*, 27:27–33, 1978.

[3] J. Felsenstein. *Inferring phylogenies*. Sinauer associates, Inc., Sunderland, MA, 2004.

[4] W. M. Fitch. Toward defining the course of evolution: Minimum change for a specified tree topology. *Syst Zool*, 20:406–416, 1971.

[5] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees: a method based on mutation distances as estimated from cytochrome c sequences is of general applicability. *Science*, 155:279–284, 1967.

[6] W. S. Fitch. Distinguishing homologous from analogous proteins. *Syst. Zool.*, 19:99–113, 1970.

[7] P. H. Harvey, A. J. Leigh Brown, John Maynard Smith, and S. Nee. *New Uses for New Phylogenies*. Oxford Univ Press, Oxford. England, 1996.

[8] M. Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, London, 1983.


[9] W.-S. Li. *Molecular evolution.* Sinauer Associates, Inc., Sunderland, MA, 1997.

[10] C. D. Michener and R. R. Sokal. A quantitative approach to a problem of classification. *Evolution*, 11:490–499, 1957.

[11] M. Nei and S. Kumar. *Molecular evolution and phylogenetics.* Blackwell Science Ltd., Oxford, London, first edition, 1998.

[12] R. D. M. Page and E. C. Holmes. *Molecular evolution. A phylogenetic approach.* Blackwell Science Ltd., Oxford, London, first edition, 1998.

[13] M. Robinson-Rechavi and D. Huchon. RRTree: relative-rate tests between groups of sequences on a phylogenetic tree. *Bioinformatics*, 16:296–297, 2000.

[14] M. Salemi and A. M. Vandamme (ed). *The phylogenetic handbook. A practical approach to DNA and protein phylogeny.* Cambridge University Press, UK, 2003.

[15] D. Sankoff and P. Rousseau. Locating the vertixes of a Steiner tree in an arbitrary metric space. *Math. Progr.*, 9:240–276, 1975.

[16] R. R. Sokal and P. H. Sneath. *Numerical taxonomy.* W. H. Freeman, San Francisco, 1963.

[17] D. L. Swofford. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.* Sinauer Associates, Sunderland, Massachusetts, 2003.

[18] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. Phylogenetic inference. In D. M. Hillis, C. Moritz, and B. K. Mable, editors, *Molecular systematics (2nd ed.)*, pages 407–514. Sinauer Associates, Inc., Sunderland, Massachusetts, 1996.

[19] D. L. Swofford and J. Sullivan. Phylogeny inference based on parsimony and other methods using PAUP*. Theory and practice. In M. Salemi and A. M. Vandamme, editors, *The phylogenetic handbook. A practical approach to DNA and protein phylogeny*, pages 160–206. Cambridge University Press, UK, 2003.

[20] E. O. Wiley, D. Siegel-Causey, D. R. Brooks, and V. A. Funk. *The Compleat Cladist.A Primer of Phylogenetic Procedures.* The University of Kansas Museum of Natural History. Lawrence, Special Publication N°19, 1991.

[21] E. Zuckerkandl and L. Pauling. Molecules as documents of evolutionary history. *J Theor Biol*, 8:357–366, 1965.