Molecular Evolution and Phylogenetics

Cambridge University Edition II

Arbiza Leonardo & Hernán Dopazo*

Pharmacogenomics and Comparative Genomics Unit

Bioinformatics Department†

Centro de Investigación Príncipe Felipe‡

CIPF

Valencia - Spain

16 - 18 October, 2006

* hdopazo@cipf.es

† http://bioinfo.cipf.es

‡ http://www.cipf.es

1. Objectives

- This short, but intensive course, has the purpose to introduce students to **the main concepts of molecular evolution and phylogenetics analysis**:
  - Homology
  - Models of Sequence Evolution
  - Molecular Adaptation
  - Cladograms & Phylograms
  - Outgroups & Ingroups
  - Rooted & Unrooted trees
  - Phylogenetic Methods: MP, ML, Distances
- The course consists of a series of **lectures and PC. Lab. sessions** that will familiarize the student with the statistical problem of phylogenetic reconstruction and its multiple uses in biology.

2. Introduction

2.1. 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!!!

2.2. Applications of phylogenies

Phylogenetic information is used in different areas of biology. From population genetics to macroevolutionary studies, from epidemiology to animal behaviour, from forensic practice to conservation ecology [1]. In spite of this broad range of applications, **phylogenies are used by making inferences from**:

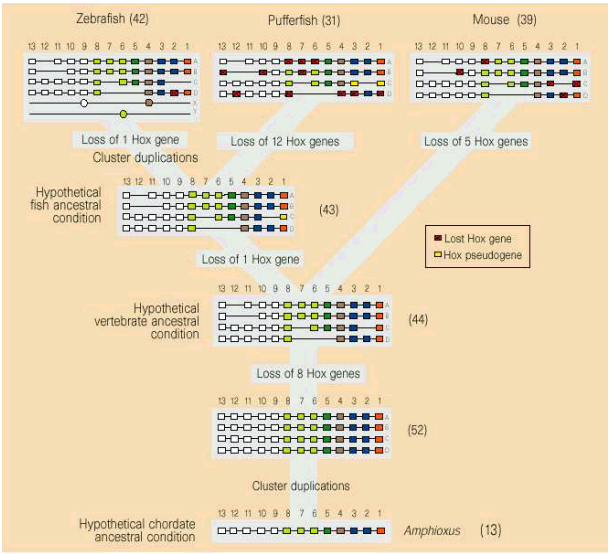
1. **Tree topology and branch lengths:**
    - Applications in **evolutionary genetics** deducing partial internal duplication of genes [26], recombination [24], reassortment [7], gene conversion [85], translocations [56] or xenology [92, 83].
    - Applications in **population genetics** in order to quantify parameters and processes like gene flow [95], mutation rate, population size [21], natural selection [30] and speciation [44] [2]
    - Applications by **estimating rates and dates** in order to check clocklike behaviour of genes [27], to date events in epidemiological studies [111], or macroevolutionary events [55, 39, 38].
    - Applications by testing **evolutionary processes** like coevolution [34], cospeciation [76, 75], biogeography [99, 33], molecular adaptation, neutrality, convergence, tissue tropisms (HIV clones), the origin of geneteic code, stress effects in bacteria, etc.

[1] See [36] for a comprehensive revision on the issue
[2] See [16] for a review on these methods.

- Applications in conservation biology [70], forensic or legal cases [45], *the list is far less than exhaustive!!!*

2. **Mapping character states on to the tree:**

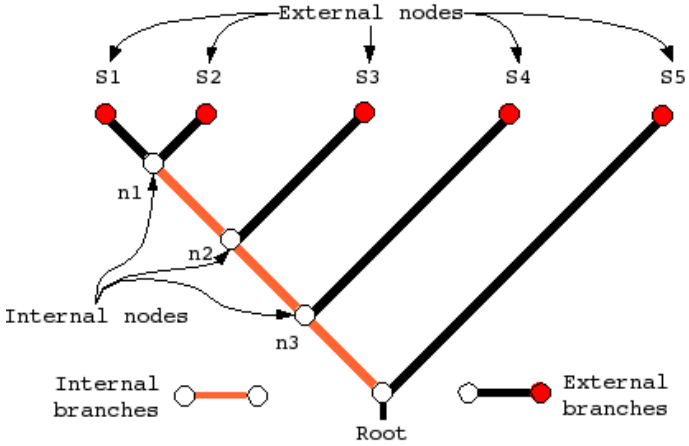
   - Applications in comparative biology [37, 5, 76], in areas like animal behaviour [64, 5], development [67], speciation and adaptation [5]



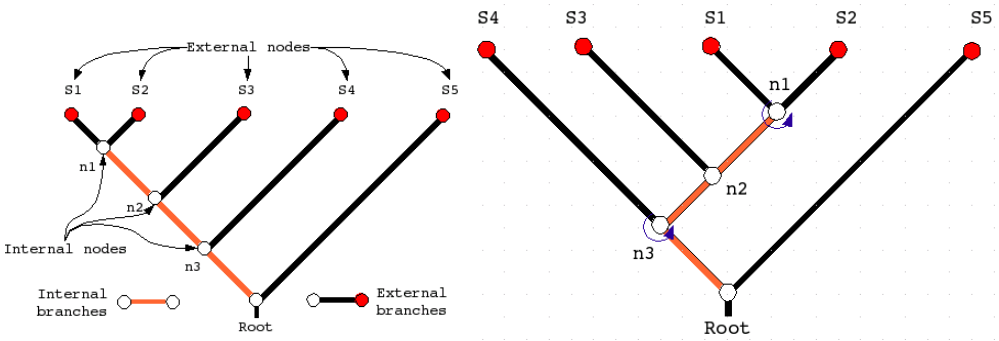
3. Tree Terminology

3.1. 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.

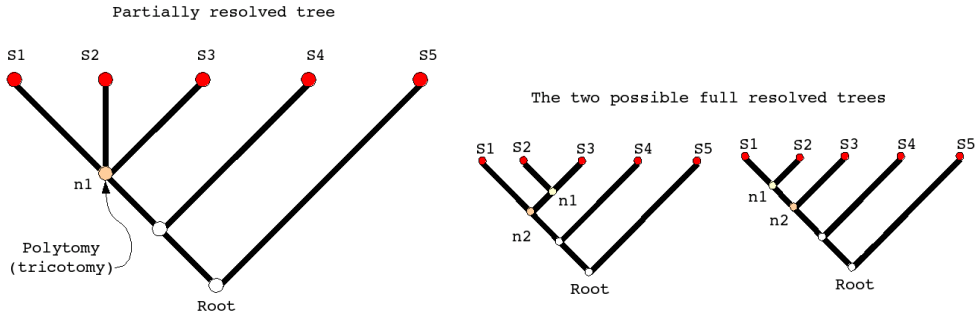


- **Root** is the common ancestor of all the sequences.
- **Topology** represents the branching pattern. Branches **can rotate** on internal nodes. Instead of the singular aspect, the folowing trees represent a single phylogeny.



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.

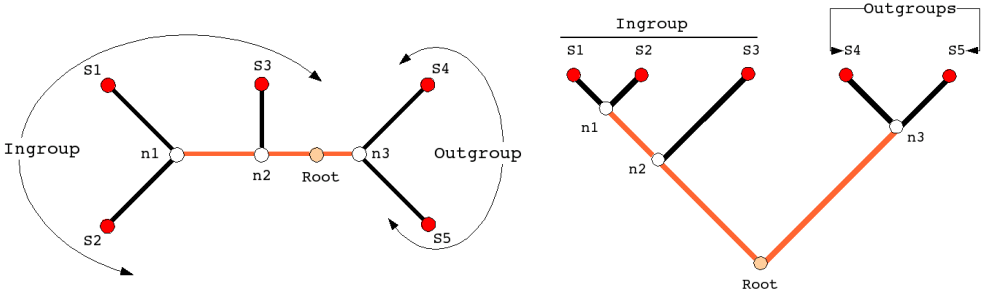


Polytomies can be solved by using more sequences, more characters or both!!!

3.2. 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**).

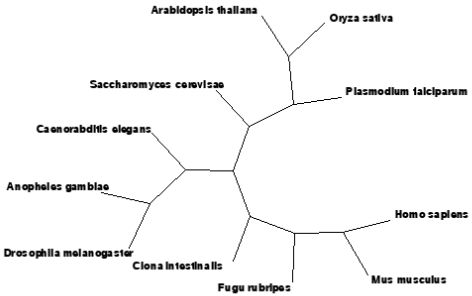


- **Unrooted trees** show the topological relationships among sequences althoug it is impossible to deduce wether nodes (n_i) represent a primitive or derived evolutionary condition.
- **Rooted trees** show the evolutionary basal and derived evolutionary relationships among sequences.

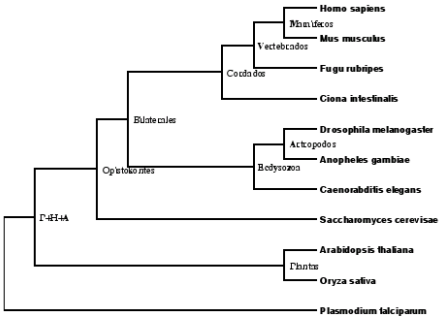
Rooting by outgroup is frequent in molecular phylogenetics!!

3.3. Cladograms & Phylograms

Trees showing branching order exclusivelly (**cladogenesis**) are principally the interest of systematists[3] to make inferences on taxonomy[4]. Those interesting in the evolutionary processes emphasize on branch lengths information (**anagenesis**).



Unrooted dendogram showing branching order

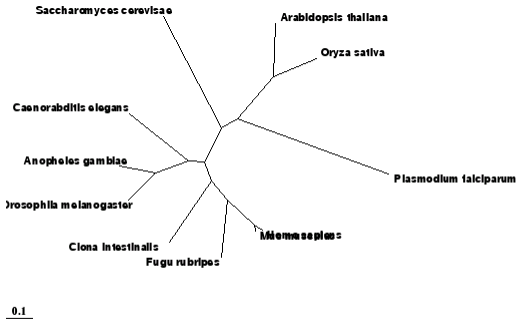


Rooted cladogram (cladistic methods)

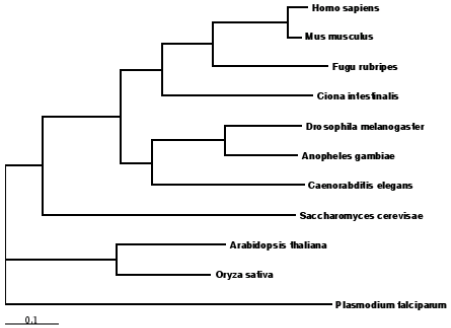
- **Dendrogram** is a branching diagram in the form of a tree used to depict degrees of relationship or resemblance.
- **Cladogram** is a branching diagram depicting the hierarchical arrangement of taxa defined by cladistic methods (the distribution of shared derived characters -synapomorphies-).

[3] The study of biological diversity.
[4] 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



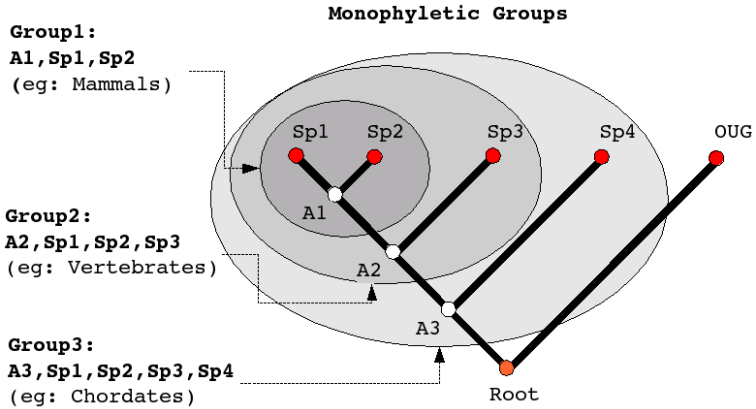
Unrooted phylogram

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

3.4. Monophyletic Groups

Taxonomic groups, to be real, must represent **a community of organisms descending from a common ancestor**.This is part of the Darwinian legacy.

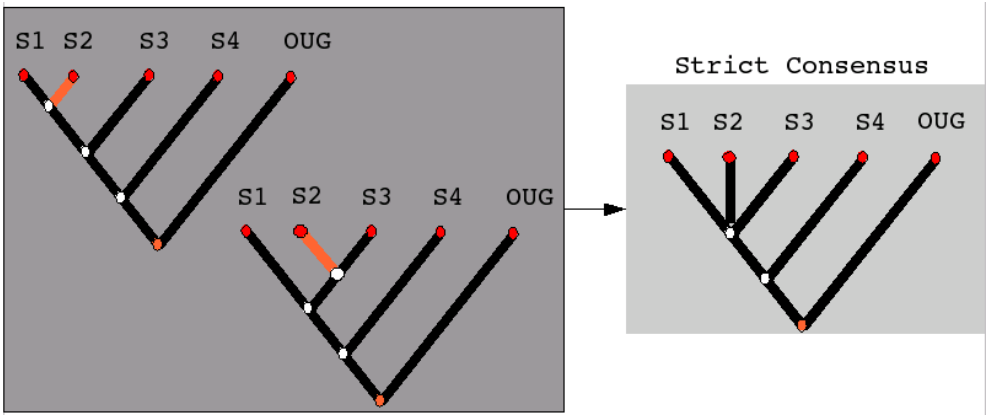
Monophyletic group represents a group of organisms with the same taxonomic title (say genus, family, phylum, etc.) that are shown phylogenetically to share a common ancestor that is exclusive to these organisms. They are, by definition, natural groups or **clades**.



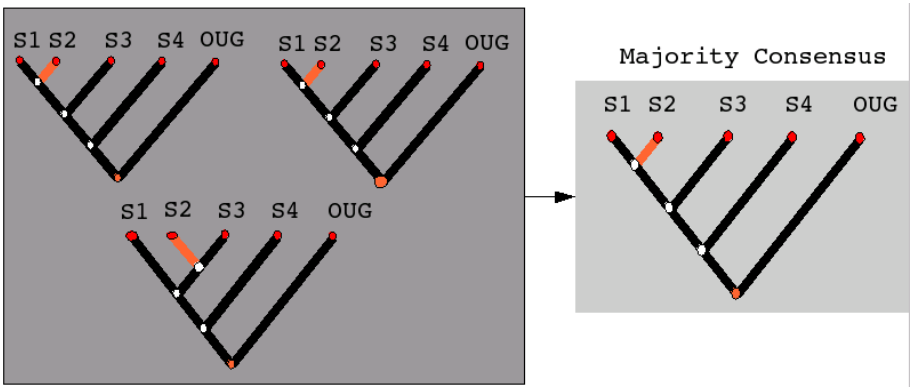
3.5. Consensus trees

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!!.[5]

A helpfull manual covering these and other concepts of the section can be obtained in [106, 77].

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

4. Homology

Richard Owen's (1847) most famous contributions to theorethical comparative anatomy were to distinguish between **homologous** and **analogous** features in organisms and to present the concept of **archetype**. The vertebrate archetype consists of a linear series of "vertebrae" and "apendages", little modified from a single basic plan. Each vertebra of the archetype is a **serial homologue** of every other vertebra of the archetype. Two corresponding vertebrae, each from different animal, are **special homologues** of one another, and **general homologues** of the corresponding vertebra of the archetype[6].



Homologue..."The same organ in different animals under every variety of form and function".
Analogue..."A part or organ in one animal which has the same function as another part or organ in a different animal".

[6] See [79] and chapters of the referenced book for a complete discussion of the term

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?

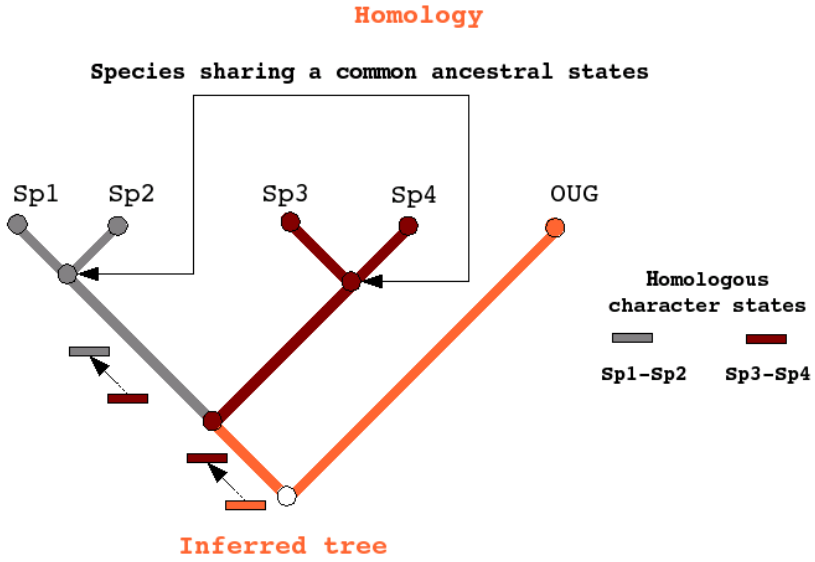
How inexplicable are the cases of serial homologies on the ordinary view of creation!

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?

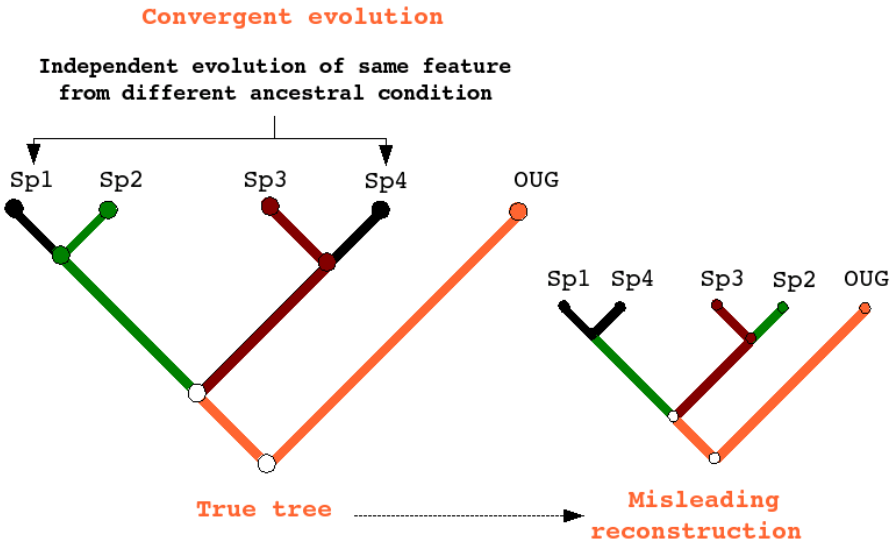
Since Darwin homology was the result of descent with modification from a common ancestor.

4.1. Homology and 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.

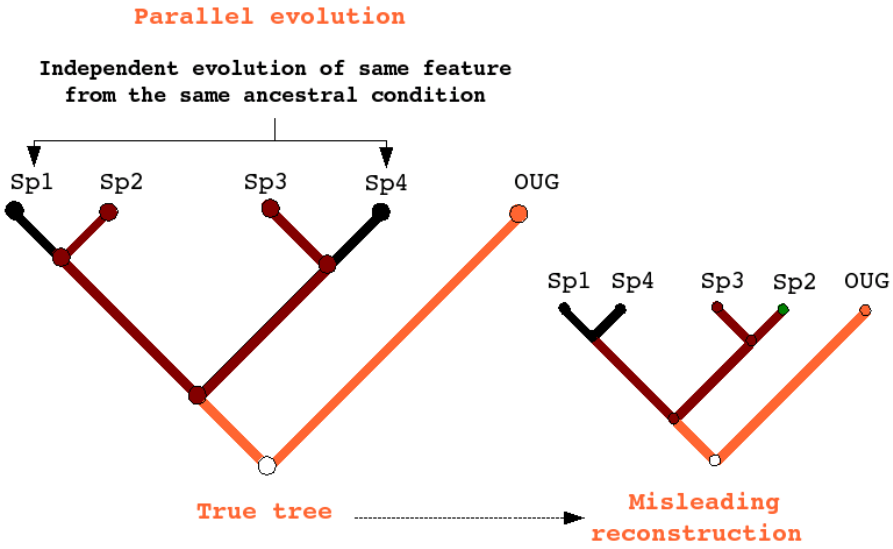


- **Convergences** are ...



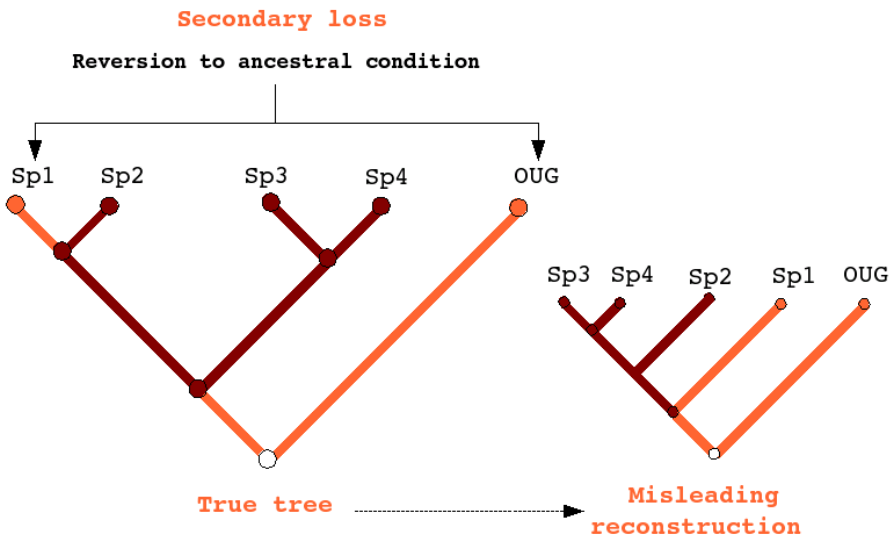
Homoplasy can provide misleading evidence of phylogenetic relationships!! (if mistakenly interpreted as homology).

- **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).

4.2. 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.

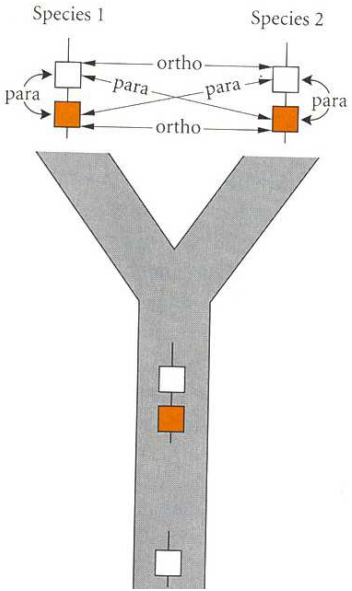
4.3. Sequence Homology

Homologous Genes are sequences that are descendant from a common ancestor (e.g., all globins).

Fitch distinguished different kinds of homologous genes [29];

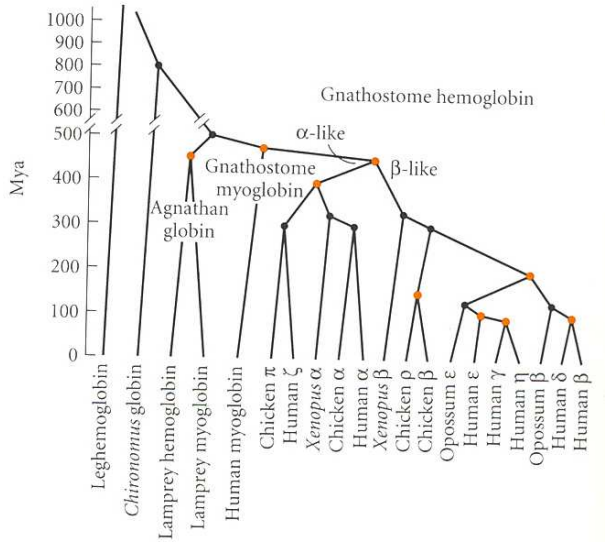
- **Ortholog**: Homologous genes that have diverged from each other after speciation events (e.g., human β- and chimp β-globin).
- **Paralog**: Homologous genes that have diverged from each other after gene duplication events (e.g., β- and γ-globin)
- **Xenolog**: Homologous genes that have diverged from each other after lateral gene transfer events (e.g., antibiotic resistance genes in bacteria).

Orthologous and Paralogous Relationships



Orthologous, Paralogous and *Xenologous* genes are *a posteriori* phylogenetic tree reconstruction definitions !!

Globins Gene Tree



4.4. Positional homology

The common ancestry of specific amino acid or nucleotide positions in different genes or sequences.

```
    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
```

5. Molecular Evolution

5.1. Molecular clock & Evolutionary Rates

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 [113].

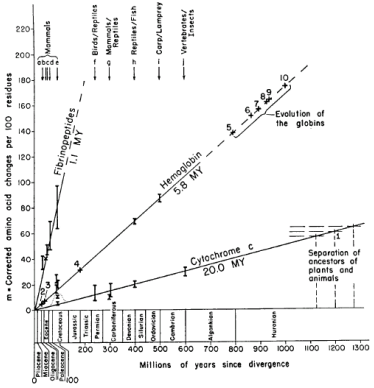


Fig. 8.3. Rates of amino acid substitution in the fibrinopeptides, hemoglobin, and cytochrome *c*. Comparisons for which no adequate time coordinate is available are indicated by numbered crosses. Point 1 represents a date of 1200 ± 75 MY (million years) for the separation of plants and animals, based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

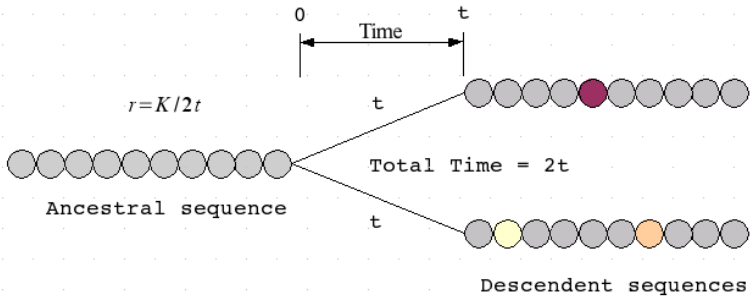
This hypothesis has estimulated much interest in the use of macromolecules

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 α- and β-chains (Kimura and Ohta, 1973b).

Region	α-chain	β-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),

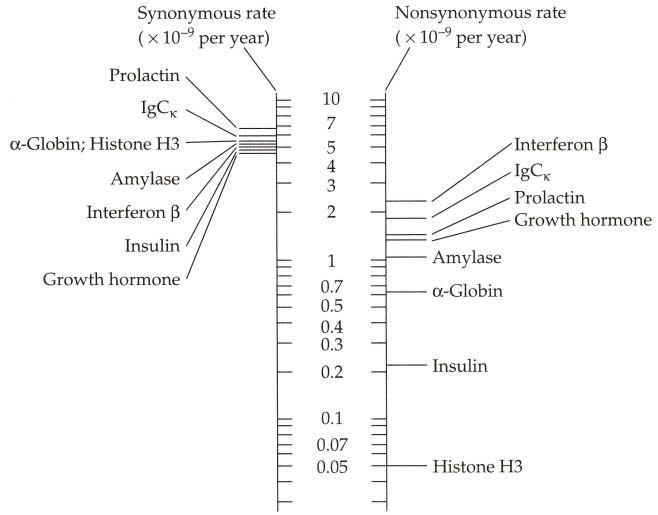


Figure 8.14 Comparison of rates of synonymous and nonsynonymous nucleotide substitutions. Synonymous rates are generally much faster and much more uniform than nonsynonymous rates. (From Kimura 1986.)

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

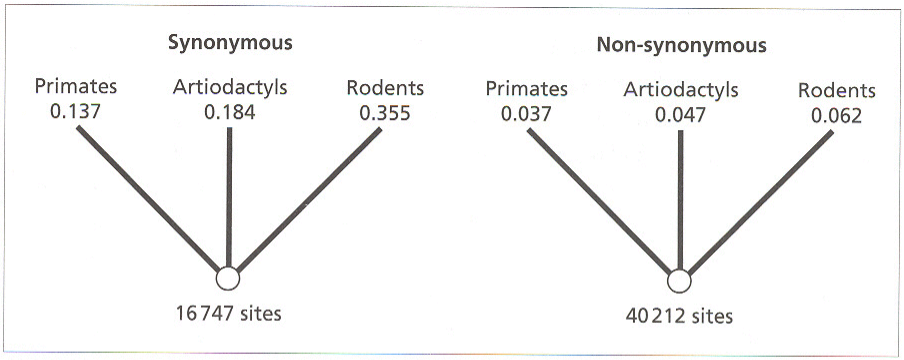
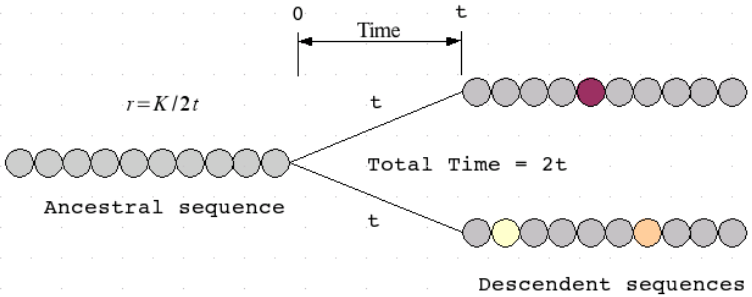


Fig. 7.14 Numbers of synonymous and non-synonymous substitutions for 49 genes from three mammalian orders: primates, rodents and artiodactyls, the phylogenetic relationships of which approximate a 'star phylogeny'. Note that, in both cases, rodents have accumulated more substitutions than primates or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

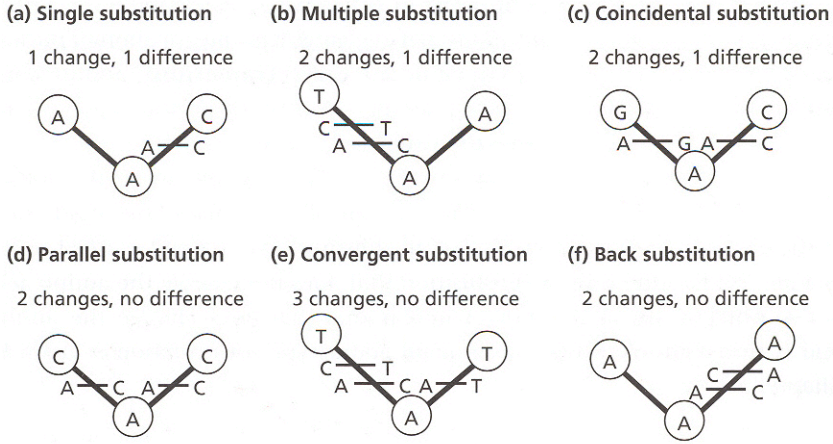
6. Evolutionary Models

6.1. Multiple Hits

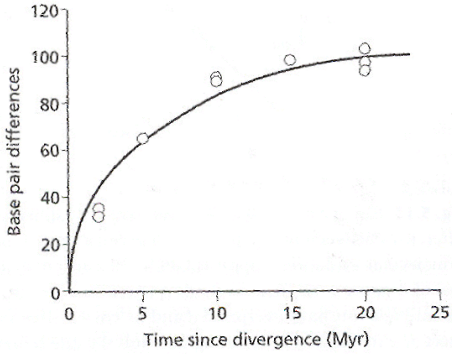
- The mutational change of DNA sequences varies with region. Even considering protein coding sequence alone, the patterns of nucleotide substitution at the first, second or third codon position are not the same.
- When two DNA sequences are derived from a common ancestral sequence, the descendant sequences 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.

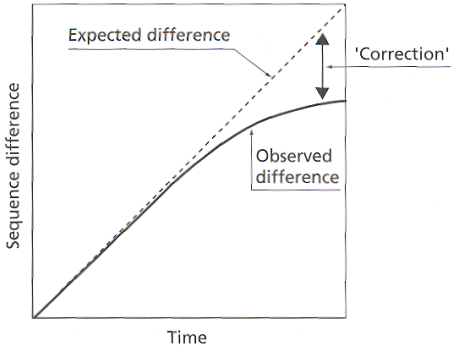


- When p is large, it gives an underestimate of the number of of substitutions, because it does not take into account **multiple substitutions**.



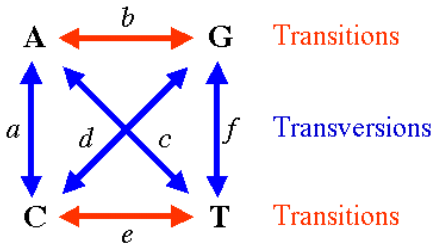
- Sequences may saturate due to multiple changes (**hits**) at the same position after lineage splitting.
- In the worst case, data may become random and all the **phylogenetic information** about relationships can be lost!!!





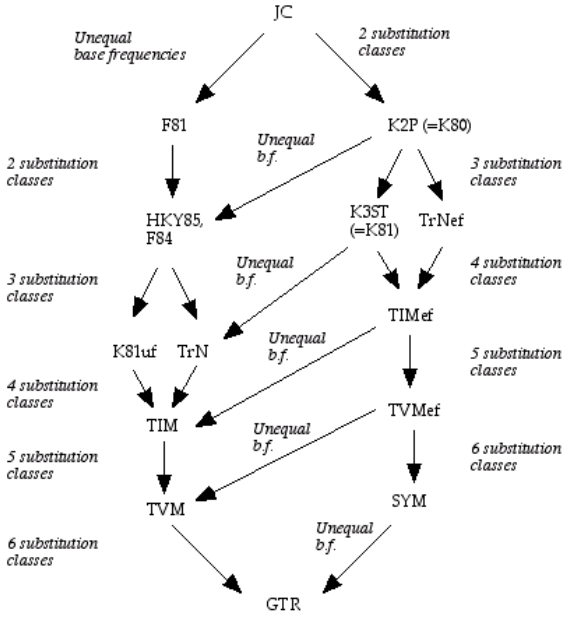
6.2. Models of nucleotide substitution

- 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.

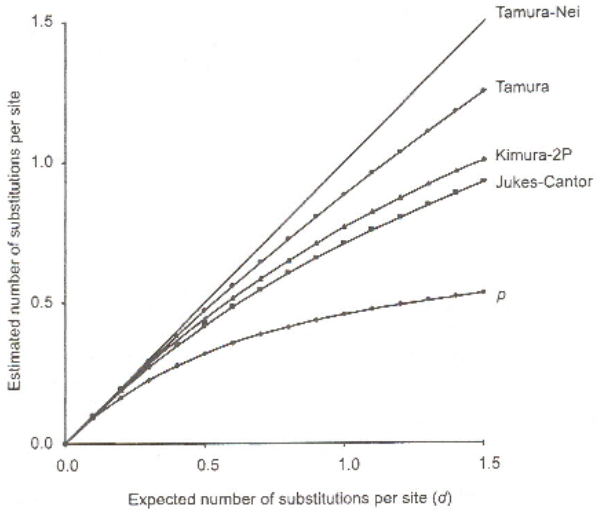


FIGURE 3.1. Estimates of the number of nucleotide substitutions obtained by different distance measures when actual nucleotide substitution follows the Tamura-Nei model. The nucleotide frequencies assumed are $g_A = 0.3$, $g_T = 0.4$, $g_C = 0.2$, and $g_G = 0.1$; and the two transition/transversion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

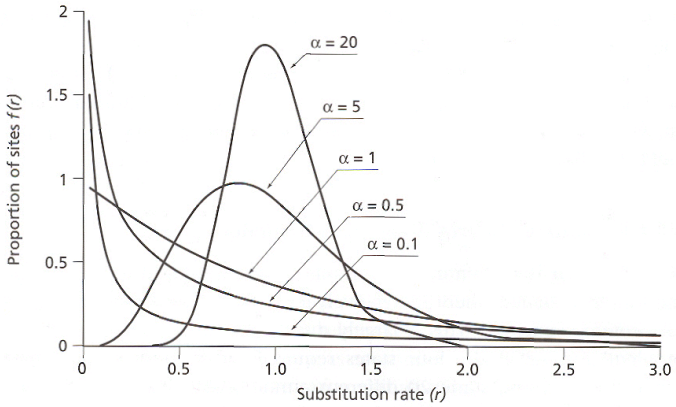
Distance correction methods share several assumptions:

- 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.

6.3. Rate heterogeneity correction

- In the evolutionary models considered, the rate of nucleotide substitution is assumed to be the same for all nucleotide. This rarely holds, and rates varies from site to site.
- In the case of protein coding genes this is obvious: 1, 2 and 3 positions.
- In the case of RNA coding genes, secondary structure consisting in loops and stems have different substitutions rates.



- Statistical analyses have suggested that the rate variation approximately follows the gamma (Γ) distribution

- Rate variation on different genes,

Type of sequences	α
Nuclear genes	
Albumin genes	1.05
Insulin genes	0.40
c-myc genes	0.47
Prolactin genes	1.37
16S-like rRNAs, stem region	0.29
16S-like rRNAs, loop region	0.58
ψη-globin pseudogenes	0.66
Viral genes	
Hepatitis B virus genomes	0.26
Mitochondrial genes	
12S rRNAs	0.16
Position 1 of four genes	0.18
Position 2 of four genes	0.08
Position 3 of four genes	1.58
D-loop region	0.17
Cytochrome *b*	0.44

- Low α values corresponds to large rate variation. As α gets larger the rate of variation diminishes, until as α approaches ∞ all sites have the same substitution rate [107].
- Models are labeled as **JC**+Γ, **K80**+Γ, **HKY**+Γ, *etc.*
- Indeed models can be corrected by considering the **proportion of invariable sites** (I) and the **nucleotide frequency** (F): (**JC**+Γ+$I+F$) ; (**K80**+Γ+$I+F$) ; (**HKY**+Γ+$I+F$); *etc.*

6.4. Selecting models of evolution

The best-fit model of evolution for a particular data set can be selected through statistical testing. The fit to the data of different models can be contrasted through **likelihood ratio tests (LRTs)** , the **Akaike (AIC)** or the **Bayesian (BIC)** information criteria[82].

A natural way of comparing two models is to contrast their likelihood using the LRT statistic:

$$\Delta = 2(log_e L_1 - log_e L_0)$$

Where L_1 is the maximum likelihood under the more parameter-rich, complex model(i.e., alternative hypothesis) and L_0 is the maximum likelihood under the less parameter-rich, simple model (i.e., null hypothesis).

When model comparison is not nested, the **AIC** criteria, which measures the expected distance between the true model and the estimated model can be used.

$$AIC_i = -2(log_e L_i + 2N_i)$$

Where N_i is the number of free parameters in the ith model and L_i is the maximum likelihood value of the data under the ith model.[7]

When LRT is significant ($p \leq 0.05$, Chi-square comparison, degrees of freedom equal to the difference in number of free parameters between the two models), the more complex model is favored.

[7] See [80] for a clear theorethical and practical explanation on sequence model test's methods.

Comparing 2 different **nested** models through an LRT means testing hypothesis about data. **MODELTEST** program [81] tests hierarchical LRTs in an ordered way and compute **AIC** values.

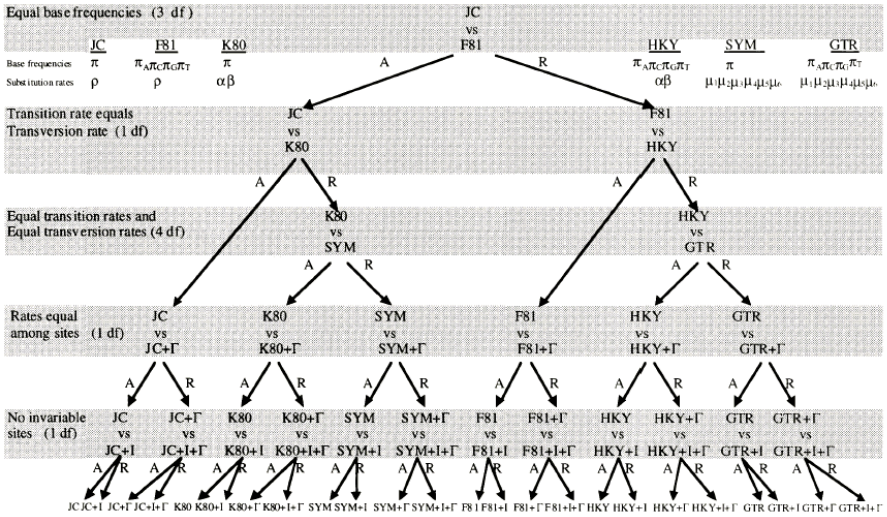


Fig. 1. Hierarchical hypothesis testing in MODELTEST. At each level the null hypothesis (upper model) is either accepted (A) or rejected (R). The models of DNA substitution are: JC (Jukes and Cantor, 1969), K80 (Kimura, 1980), SYM (Zharkikh, 1994), F81 (Felsenstein, 1981), HKY (Hasegawa *et al.*, 1985), and GTR (Rodríguez *et al.*, 1990). Γ: shape parameter of the gamma distribution; I: proportion of invariable sites. df: degrees of freedom. 1: equal base frequencies (0.25), π_A: frequency of adenine, π_C: frequency of cytosine, π_G: frequency of guanine, π_T: frequency of thymine. ρ: equal substitution rate, α: transition rate, β: transversion rate; μ_1: A⇒C rate, μ_2: A⇒G rate, μ_3: A⇒T rate, μ_4: C⇒G

6.5. Amino acid models

In contrast to DNA, the modeling of amino acid replacement has concentrated on the **empirical approach**.
Dayhoff [12] developed a model of protein evolution that resulted in the development of a set of widely used replacement matrices. In the Dayhoff approach,

- Replacement rates are derived from alignments of protein sequences 85% identical,
- This ensures that the likelihood of a particular mutation (e.g., L $\mapsto$ V) being the result of a set of successive mutations (e.g., L $\mapsto$ x $\mapsto$ y $\mapsto$ V) is low.
- An implicit instantaneous rate matrix is estimated, and replacement probability matrices $\mathbf{P}(T)$ are generated at different values of T
- One of the main uses of the Dayhoff matrices has been in databases search methods, PAM50, PAM100, PAM250 corresponding to $\mathbf{P}(0.5)$, $\mathbf{P}(1)$ and $\mathbf{P}(2.5)$, respectivelly.
- The number 250 in PAM250 corresponds to an average of 250 amino acid replacements per 100 residues from a data set of 71 aligned sequences.

A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z	
0.4	0.0	-0.4	0.0	0.0	-0.8	0.2	-0.2	-0.2	-0.2	-0.4	-0.2	0.0	0.2	0.0	-0.4	0.2	0.2	0.0	-1.2	-0.6	0.0	A
	0.5	-0.9	0.6	0.4	-1.0	0.1	0.3	-0.4	0.1	-0.7	-0.5	0.4	-0.2	0.3	-0.1	0.1	0.0	-0.4	-1.1	-0.6	0.4	B
		2.4	-1.0	-1.0	-0.8	-0.6	-0.6	-0.4	-1.0	-1.2	-1.0	-0.8	-0.6	-1.0	-0.8	0.0	-0.4	-0.4	-1.6	0.0	-1.0	C
			0.8	0.6	-1.2	0.2	0.2	-0.4	0.0	-0.8	-0.6	0.4	-0.2	0.4	-0.2	0.0	0.0	-0.4	-1.4	-0.8	0.5	D
				0.8	-1.0	0.0	0.2	-0.4	0.0	-0.6	-0.4	0.2	-0.2	0.4	-0.2	0.0	0.0	-0.4	-1.4	-0.8	0.6	E
					1.8	-1.0	-0.4	0.2	-1.0	0.4	0.0	-0.8	-1.0	-1.0	-0.8	-0.6	-0.6	-0.2	0.0	1.4	-1.0	F
						1.0	-0.4	-0.6	-0.4	-0.8	-0.6	0.0	-0.2	-0.2	-0.6	0.2	0.0	-0.2	-1.4	-1.0	-0.1	G
							1.2	-0.4	0.0	-0.4	-0.4	0.4	0.0	0.6	0.4	-0.2	-0.2	-0.4	-0.6	0.0	-0.4	H
								1.0	-0.4	0.4	0.4	-0.4	-0.4	-0.4	-0.4	-0.2	0.0	0.8	-1.0	-0.2	-0.4	I
									1.0	-0.6	0.0	0.2	-0.2	0.2	0.6	0.0	0.0	-0.4	-0.6	-0.8	0.1	K
										1.2	0.8	-0.6	-0.6	-0.4	-0.6	-0.6	-0.4	0.4	-0.4	-0.2	-0.5	L
											1.2	-0.4	-0.4	-0.2	0.0	-0.4	-0.2	0.4	-0.8	-0.4	-0.3	M
												0.4	-0.2	0.2	0.0	0.2	0.0	-0.4	-0.8	-0.4	0.2	N
													1.2	0.0	0.0	0.2	0.0	-0.2	-1.2	-1.0	-0.1	P
														0.8	0.2	-0.2	-0.2	-0.4	-1.0	-0.8	0.6	Q
															1.2	0.0	-0.2	-0.4	0.4	-0.8	0.6	R
																0.4	0.2	-0.2	-0.4	-0.6	-0.1	S
																	0.6	0.0	-1.0	-0.6	-0.1	T
																		0.8	-1.2	-0.4	-0.4	V
																			3.4	0.0	-1.2	W
																				2.0	-0.8	Y
																					0.6	Z

Several later groups have attempted to extend Dayhoff's methodology or re-apply her analysis using later databases with more examples.

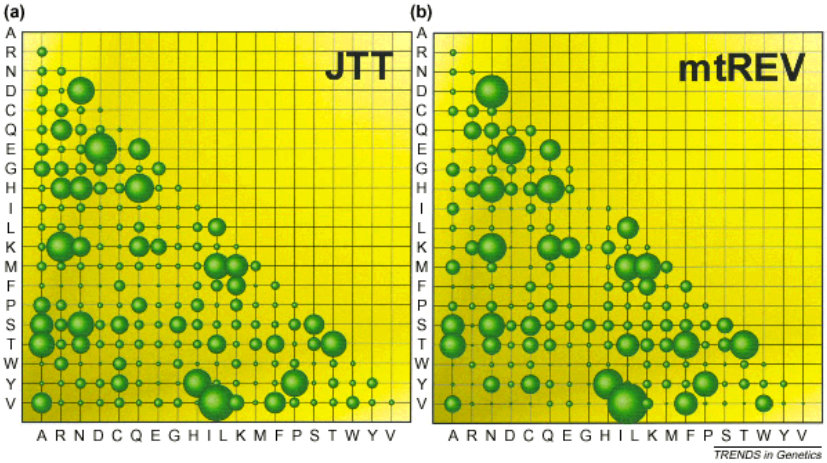
- Jones, et al. [49] used the same methodology as Dayhoff but with modern databases and for membrane spanning proteins.

The BLOSUM series of matrices were created by Henikoff [41]. Features,

- Derived from local, ungapped alignments of distantly related sequences,
- All matrices are directly calculated; no extrapolations are used,
- The number of the matrix (BLOSUM62) refers to the minimum % identity

of the blocks used to build the matrix; greater numbers, lesser distances,

- The BLOSUM series of matrices generally perform better than PAM matrices for local similarity searches.
- Specific matrices modeling mitochondrial proteins exists [1, 63]
- Indeed, others approaches to have recently been done [62, 71, 104][8]



[8] See [61, 105] for a review of evolutionary sequence models

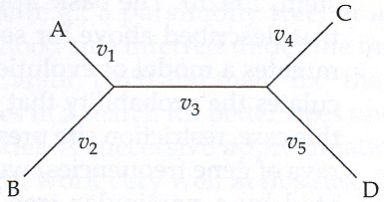
7. Distance Methods

Distance matrix methods is a major family of phylogenetic methods trying to fit a tree to a matrix of pairwise distance [10, 28]. 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.

7.1. Ultrametric & Additive Trees

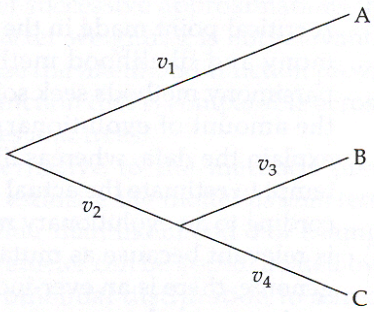
(A)



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:

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

7.2. Cluster Analysis

Cluster analysis derived from clustering algorithms popularized by Sokal and Sneath[97]

7.2.1. 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 [68].

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.

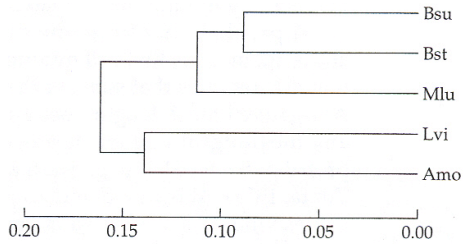
   The variants of UPGMA are in the step 4. Weighted PGMA(WPGM::$d_{ku} = d_{ki} + d_{kj}/2$). Complete linkage ($d_{ku} = max(d_{ki}, d_{kj})$. Single linkage($d_{ku} = min(d_{ki}, d_{kj})$.

	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.

7.2.2. 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 $\mapsto r_i = \sum_{k=1}^{N} d_{ik}$ [9].
2. Create a rate-corrected distance matrix $(\mathbf{M})$ in which the elements are defined by $\mapsto M_{ij} = d_{ij} - (r_i + r_j)/(N-2)$ [10].
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) $\mapsto v_{ij} = d_{ij}$

[9] N is the number of terminal nodes

[10] 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

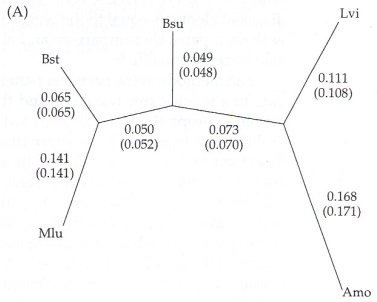
Bsu to node 2 distance = 0.1222/2 + (0.2631 – 0.2869)/2 = 0.0492
node 1 to node 2 distance = 0.1222 – 0.0492 = 0.0730

	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.

7.3. 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

8. 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 τ 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).[11]

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

Determining the length of the tree is computed by algorithmic methods[25, 90]. However, we will show how to calculate the length of a particular tree topology ((W,Y),(X,Z))[12] for a specific site of a sequence, using Fitch (A) and transversion parsimony (B)[13]:

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

(A)

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

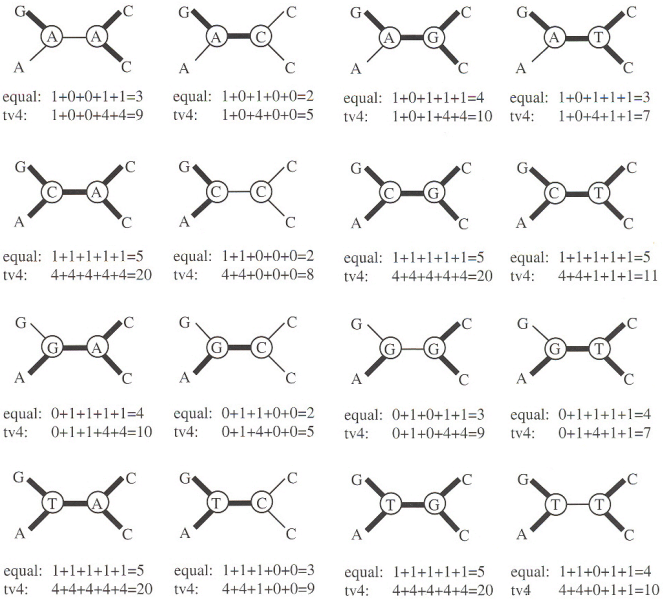
(B)

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

- 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**,[14]
- 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,

[12] Newick format
[13] Matrix character states: A,C,G,T
[14] 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,



8.1. 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.

9. Searching Trees

9.1. 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 [19] 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	$> 6\text{x}10^{81}$	$> 2\text{x}10^{76}$

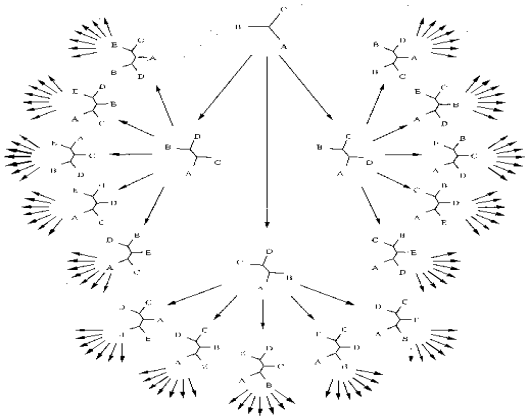
The observable universe has about $8.8\text{x}10^{77}$ atoms

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!!

9.2. 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,



9.3. 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

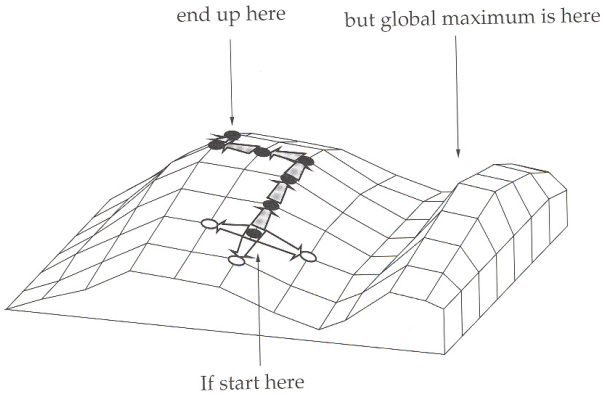
9.3.1. Greedy Algorithms

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[15]

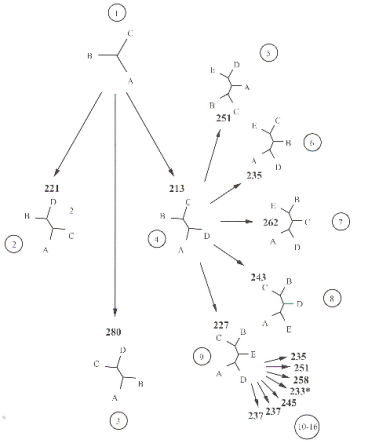
Both algoritms are prone to entrapment in local optima

[15] See Additional Material



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.

9.3.2. 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.*

Tree Bisection & Reconnection

- Divide tree into two parts,
- Reconnect by a pair of branches, attempting every possible pair of branches to rejoin
- NNI and SPR are subsets of TBR

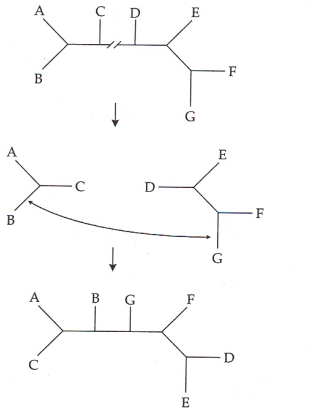


Figure 28 Branch swapping by tree bisection and reconnection. The tree is bisected along a branch, yielding two disjoint subtrees. The subtrees are then reconnected by joining a pair of branches, one from each subtree. All possible bisections and pairwise reconnections are evaluated.

10. Statistical Methods

10.1. Maximum Likelihood

♣ The phylogenetic methods described **infered the history** (*or the set of histories*) that were **most consistent with a set of observed data**.
All the methods explained used **sequences as data** and give one or more **trees as phylogenetic hypotheses**. Then, they use the logic of:

$$P(H/D)$$

♠ **Maximum Likelihood** (*ML*)[16] methods (*or maximum probability*) computes **the probability of obtaining the data** (*the observed aligned sequences*) **given a defined hypothesis** (*the tree and the model of evolution*). That is:

$$P(D/H)$$

A coin example

The ML estimation of the heads probabilities of a coin that is tossed n times.

[16] ML was invented by Ronal A. Fisher [23]. Likelihood methods for phylogenies were introduced by Edwars and Cavalli-Sforza for gene frequency data [9]. Felsenstein showed how to compute ML for DNA sequences [20].

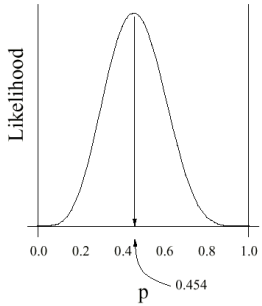
If tosses are all **independent**, and all have the same **unknown heads probability** p, then the observing sequence of tosses:

HHTTHTHHTTT

we can calculate the ML of these data as:

$$L = Prob(D/p) = pp(1-p)(1-p)p(1-p)pp(1-p)(1-p)(1-p) = p^5(1-p)^6$$

Ploting L against p, we observe the probabilities of the same data (D) for different values of p.



Thus the ML or the maximum probability to observe the above sequence of events is at $p = 0.4545$,

That is: $\frac{5}{11} \Rightarrow (\frac{heads}{heads+tails})$

⋆ This **can be verified** by taking the derivative of L with respect to p:

$$\frac{dL}{dp} = 5p^4(1-p)^6 - 6p^5(1-p)^5$$

equating it to zero, and solving:

$$\frac{dL}{dp} = p^4(1-p)^5[5(1-p) - 6p] = 0 \longrightarrow \hat{p} = 5/11$$

⋆ More easily, likelihoods are often maximized **by maximizing their logarithms**:

$$lnL = 5lnp + 6ln(1-p)$$

whose derivative is:

$$\frac{d(lnL)}{dp} = \frac{5}{p} - \frac{6}{1-p} = 0 \longrightarrow \hat{p} = 5/11$$

The likelihood of a sequence

Suppose we have:

- **Data**: a sequence of 10 nucleotides long, say **AAAAAAAATG**
- **Model**: Jukes-Cantor $\longrightarrow f_{(A,C,G,T)} = \frac{1}{4}$
- **Model**: $Model_1 \longrightarrow f_{(A,C,G,T)} = \frac{1}{2}; \frac{1}{5}; \frac{1}{5}; \frac{1}{10}$

$$L_{JC} = (\tfrac{1}{4})^8.(\tfrac{1}{4})^0.(\tfrac{1}{4}).(\tfrac{1}{4}) = (\tfrac{1}{4})^{10} = 9.53\text{x}10^{-07}$$

$$L_{M_1} = (\tfrac{1}{2})^8.(\tfrac{1}{5})^0.(\tfrac{1}{5}).(\tfrac{1}{10}) = 7.81\text{x}10^{-05}$$

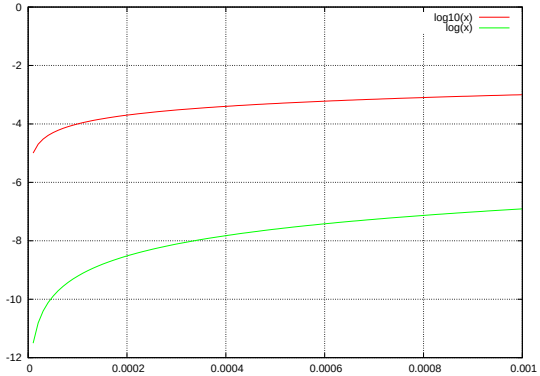
L_{M_1} is almost 100 times higher than to L_{JC} model

Thus the JC model is not the best model to explain this data

Since likelihoods takes the form of:

$$\prod_{i=1}^{n} = x_i \text{ , where: } 0 \leq x_i \leq 1 \text{ and generally } n \text{ is large}$$

it is convenient **to report ML results as** lnL **or** $log_{(10)}L$



$$lnL_{(JC)} = -14.2711 \text{ ; } lnL_{(M_1)} = -9.4575$$

When the more positive (*less negative* lnL *values*) **the best likelihood**

The likelihood of a one-branch tree

Suppose we have:

- **Data**:
  - Sequence 1 : 1 nucleotide long, say **A**
  - Sequence 2 : 1 nucleotide long, say **C**
  - Sequences are related by the simplest tree: **a single branch**
- **Model**:
  - Jukes-Cantor $\longrightarrow f_{(A,C,G,T)} = \frac{1}{4}$
  - $\mathbf{A} \overset{p}{\longleftrightarrow} \mathbf{C}$; $p = 0.4$

$$\text{So, } L_{tree} = \tfrac{1}{4}.(0.4) = 0.1$$

Since the model is **reversible**:

$$L_{tree:A \to C} = L_{tree:C \to A}$$

Real Models

Suppose we have:

- **Data**:

Sequence 1 **C C A T**

Sequence 2 **C C G T**

- **Model**:[17]

$$\pi = [0.1, 0.4, 0.2, 0.3]$$

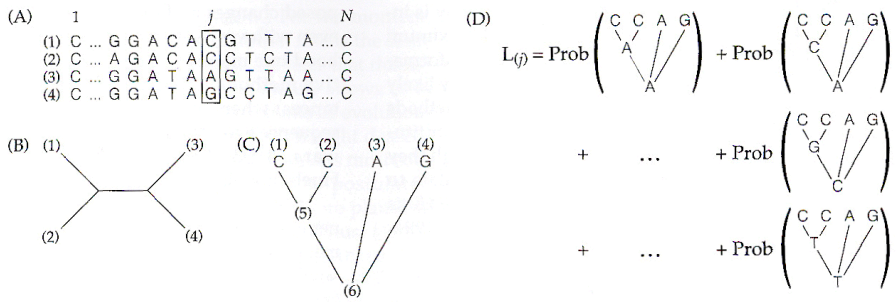
$$\boldsymbol{P} = \begin{bmatrix} 0.976 & 0.01 & 0.007 & 0.007 \\ 0.002 & 0.983 & 0.005 & 0.01 \\ 0.003 & 0.01 & 0.979 & 0.007 \\ 0.002 & 0.013 & 0.005 & 0.979 \end{bmatrix}$$

$$L_{(Seq._1 \rightarrow Seq._2)} = \pi_C P_{C \rightarrow C} \pi_C P_{C \rightarrow C} \pi_A P_{A \rightarrow G} \pi_T P_{T \rightarrow T}$$
$$0.4\text{x}0.983\text{x}0.4\text{x}0.983\text{x}0.1\text{x}0.007\text{x}0.3\text{x}0.979$$
$$= 0.0000300$$

$$lnL_{tree:Seq_1 \rightarrow Seq_2} = -10.414$$

[17] Note that the base composition sum one, but indeed the the rows of substitution matrix sum one. Why?

L computation in a real problem



- Tree after rooting in an arbitrary node (reversible model).
- The likelihood for a particular site is the sum of the probabilities of every possible reconstruction of ancestral states given some model of base substitution.
- The likelihood of the tree is the product of the likelihood at each site.

$$L = L_{(1)} \cdot L_{(2)} \cdot ... \cdot L_{(N)} = \prod_{j=1}^{N} L_{(j)}$$

- The likelihood is reported as the sum of the log likelihhod of the full tree.

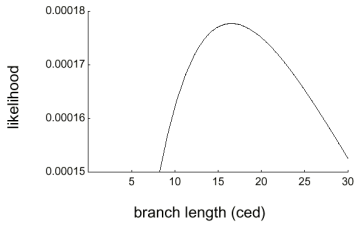
$$lnL = lnL_{(1)} + lnL_{(2)} + ... + lnL_{(N)} = \sum_{j=1}^{N} lnL_{(j)}$$

Modifying branch lengths

At moment **for L computation** we do not take into acount the posibility of different branch lengths. However, we can infer that:

- For very short branches, the probability of characters staying the same is high and the probability of it changing is low.
- For longer branches, the probability of character change becomes higher and the probability of staying the same is low
- Previous calculations are based on a Certain Evolutionary Distance (**CED**)
- We can calculate the branch length being $2, 3, 4, ...n$ times larger (**nCED**) by multiplying the substitution matrix **P** by itself n times.[18]

branch length (ced units)	likelihood
1	0.0000300
2	0.0000559
3	0.0000782
10	0.000162
15	0.000177
20	0.000175
30	0.000152

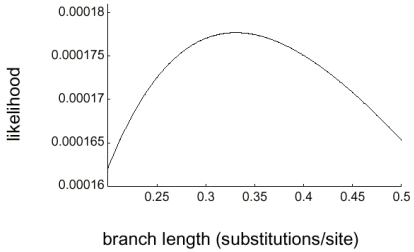


[18] At time the branch length increases, the probability values on the diagonal going down at time the prob. off the diagonal going up. Why?

Finally,

- The correct transformation of branch lengths (t) measured in *substitutions per site* is computed and maximized by:

$$P(t) = e^{Qt}$$



Where Q is the instantaneous rate matrix specifying the rate of change between pairs of nucleotides per instant of time dt.

10.2. Pros & Cons of ML

- **Pros**:
  - Each site has a likelihood,
  - Accurate branch lengths,
  - There is no need to correct for "anything",
  - The model could include: instantaneous substitution rates, estimated frequencies, among site rate variation and invariable sites,
  - If the model is correct, the tree obtained is "correct",
  - All sites are informative,
- **Cons**:
  - If the model is correct, the tree obtained is "correct",
  - Very computational intensive,

10.3. Bayesian inference

♣ **Maximum Likelihood** will find the tree that is most likely to have produced the observed sequences, or formally $P(D/H)$ (the probability of seeing the data given the hypothesis).

♠ **A Bayesian approach** will give you the tree (or set of trees) that is most likely to be explained by the sequences, or formally $P(H/D)$ (the probability of the hypothesis being correct given the data).

♢ **Bayes Theorem** provides a way to calculate the probability of a model (*tree topology and evolutionary model*) from the results it produces (*the aligned sequences we have*), what we call a **posterior probability**[19].

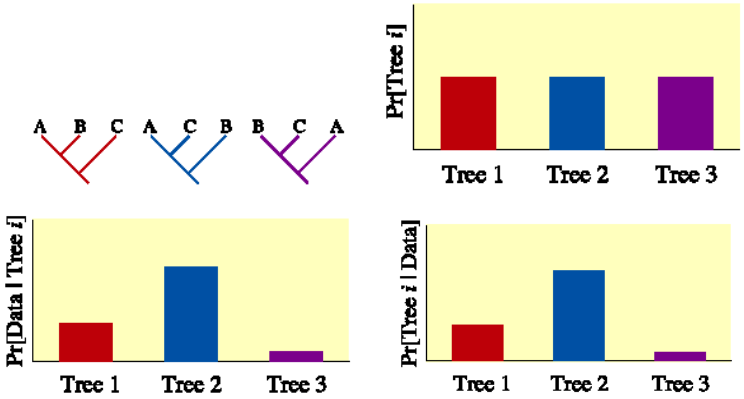
Thomas Bayes (1702-1761)

$$P(\theta/D) = \frac{P(\theta) \cdot P(D/\theta)}{P(D)}$$

[19] See [57, 47, 46] for a clear explanation on bayesian phylogenetic method.

The main components of Bayes analysis

- $P(\theta)$ The **prior probability** of a tree represents the probability of the tree before the observations have been made. Typically, all trees are considered equally probable.

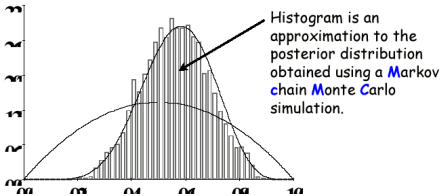


- $P(D/\theta)$ The **likelihood** is proportional to the probability of the observations (data sets) conditional on the tree.
- $P(\theta/D)$ The **posterior probability** of a tree is the probability conditional on the observations. It is obtained combined the prior and the likelihood using the Bayes' formula

How to find the solution

There's no analytical solution for a Bayesian system. However, giving:

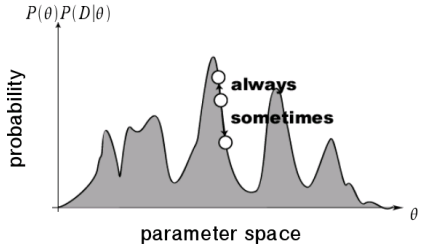
- **Data**: Sequence data,
- **Model**: The evolutionary model, base frequencies, among site rate variation parameters, a tree topology, branch lengths
- **Priors** distribution on the model parameters, and
- **A method** for calculating posterior distribution from prior distribution and data: **MCMC** technique[20]



Posterior probabilities can be estimated!!!

[20] Markov Chain Monte Carlo or the Metropolis-Hastings algorithm. See [57] for an easy explanation of the techniques.

- Each step in a Markov chain **a random modification** of the tree topology, a branch length or a parameter in the substitution model (e.g. substitution rate ratio) is assayed.
- If the **posterior computed is larger** than that of the current tree topology and parameter values, the proposed step **is taken**.
- Steps downhill are not authomatic accepted, depending on the magnitude of the decrease.



- Using these rules, the **Markov chain visits regions** of the tree space **in proportion of their posterior**.
- Suppose you sample 100,000 trees and a particular clade appears in 74,695 of the sampled trees. The probability (giving the observed data) that the group is monophyletic is 0.746, because **MC visits trees in proportion to their posterior probabilities**.

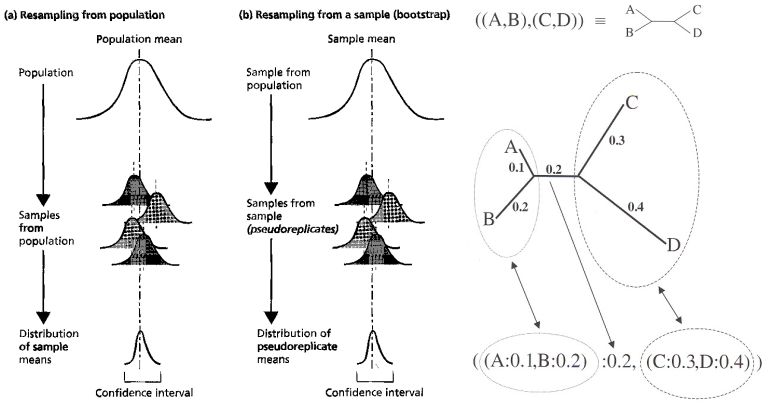
10.4. Pros & Cons of BI

- **Pros**:
  - Faster than ML,
  - Accurate branch lengths,
  - There is no need to correct for "anything",
  - The model could include: instantaneous substitution rates, estimated frequencies, among site rate variation and invariable sites,
  - If the dataset is correct, the tree obtained is "correct",
  - All sites are informative,
  - There is no neccesary bootstrap interpretations
- **Cons**:
  - To what extent is the posterior distribution influenced by the prior?
  - How do we know that the chains have converged onto the stationary distribution?
  - **A solution**: Compare independent runs starting from different points in the parameter space

11. Tree Confidence

11.1. 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.



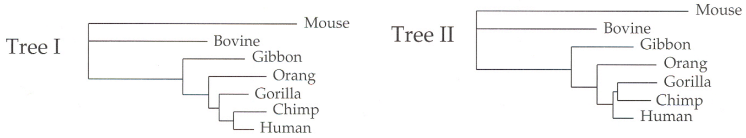
- Each sample from the original sample is a **pseudoreplicate**. By generation many hundred or thousand pseudoreplicates, a *majority consensus rule tree* can be obtained.
- High bootstrap values $> 90\%$ is indicative of strong **phylogenetic signal**.
- Bootstrap can be viewed as a way of exploring 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 require to overturn a group, then when greather the best!).[21]
- **DI** & **BPs** generally correlates!!

[21] See [102] for a practical example using PAUP*[100]

11.2. Paired site tests

The basic idea of paired sites tests is that we can compare two trees for either parsimony or likelihood or likelihood scores.

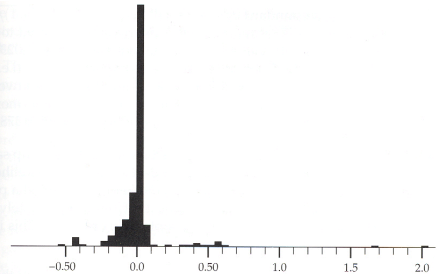
- The expected log-likelihood of a tree is the average log-likelihood we would get per site as the number of sites grows withouth limit.
- If evolution is independent, then if 2 trees have equal expected log-likelihoods, differences must be zero.
- If we do a statistical test of whether the mean of these differences is zero, we are also testing whether there is significant statistical evidence that one tree is better than another.



- The original **Kishino & Hasegawa test** (KHT) [53] calculates the z score; $z = \frac{D}{\sqrt{V_D}}$
- The z score is assumend to be normally distribuited. If z-score > 1.96, a topology is rejected at 0.05%.

- The **RELL test** (*resampling-estimated log-likelihood*) where the variance of distance log-likelihood differences is obtained by bootstrap method.

Tree	Site 1	2	3	4	5	6	...	231	232	ln L
I	−2.971	−4.483	−5.673	−5.883	−2.691	−8.003	...	−2.971	−2.691	−1405.61
II	−2.983	−4.494	−5.685	−5.898	−2.700	−7.572	...	−2.987	−2.705	−1408.80
Diff	+0.012	+0.111	+0.013	+0.015	+0.010	−0.431	...	+0.012	+0.010	+3.19



- When more than two topologies are contrasted, a multiple topology testing must be performed. **Shimodaira & Hasegawa test** (SHT) [93], **Goldman, Anderson & Rodrigo test** (SOWH) [31] and **the expected likelihood weights** method (ELW) [98] are some of the most used methods to test many alternative topologies.[22]

[22] Tree-Puzzle [91] is one of the multiple programs containing many of the tests here discussed.

12. PC Lab

12.1. Download Programs

- PHYLIP http://evolution.genetics.washington.edu/phylip.html
- PAML http://abacus.gene.ucl.ac.uk/software/paml.html
- MEGA http://www.megasoftware.net/
- TREE-PUZZLE http://www.tree-puzzle.de/
- MrBayes http://morphbank.ebc.uu.se/mrbayes/download.php
- PHYML http://atgc.lirmm.fr/phyml/
- MODELTEST http://darwin.uvigo.es/software/modeltest.html
- PROTESTS http://darwin.uvigo.es/software/prottest.html
- Hyphy http://www.hyphy.org/current/index.php
- TreeView http://taxonomy.zoology.gla.ac.uk/rod/treeview.html
- njplot http://pbil.univ-lyon1.fr/software/njplot.html

12.2. Download Data Sets

- PHYLIP format
  - ADN (HIV) http://bioinfo.ochoa.fib.es/hdopazo/download/hiv1_phy.txt
  - ADN (MtVert) http://bioinfo.ochoa.fib.es/hdopazo/download/mtv1_phy.txt
  - Proteins (GPD) http://bioinfo.ochoa.fib.es/hdopazo/download/gpd2_phy.txt
- NEXUS format
  - ADN (HIV) http://bioinfo.ochoa.fib.es/hdopazo/download/hiv1_nex.txt
  - Proteins (GPD) http://bioinfo.ochoa.fib.es/hdopazo/download/gpd2_nex.txt
- MODELTEST format
  - ADN (MtVert) http://bioinfo.ochoa.fib.es/hdopazo/download/mtv1_mdt.txt
  - Lnscores http://bioinfo.ochoa.fib.es/hdopazo/download/mtv1_modelscores.txt
- MrBayes format
  - ADN (HIV) http://bioinfo.ochoa.fib.es/hdopazo/download/hiv1_by.txt

12.3. Excercises

1. **Distance using PHYLIP**[23].

[23] Remember to put the data set in the exe' PHYLIP folder.

- Using **hiv-phy.txt** and **DNADIST** program, obtain distance matrices under JC69, F84[24] and F84+C [25] models. Compare values.
- Obtain UPGMA from JC69 and NJ trees from F84. Compare topologies.
- Using **mtv1-phy.txt**, obtain K80+Γ distances using $\alpha = 0.1, 10$. Compare values.
- Obtain NJ trees. Compare both topologies.
- Obtain LS (FM) & ME trees using **FITCH** program under F84 and JC69 models. Compare topologies.
- Define all the **monophyletic groups**.

2. **Bootstrap using PHYLIP**.

- Obtain 100 **hiv-phy.txt** randomized matrices with **SEQBOOT**.
- Obtain the corresponding LS (FM) trees using F84 model.
- Calculate BPs values using **CONSENSE** program.

3. **Parsimony & Likelihood using PHYLIP**.

- Using **hiv-phy.txt** and **DNAPARS** program, obtain MP tree/s under **Fich** optimization.

[24] **Warning**: Do not re-write outfiles!!!

[25] Where C represent categories of the 1, 2 and 3 position of the sequences evolving at 2, 1 and 20 relative rates.

- The same using **transversion** parsimony.
- Select the correct options to estimate ancestral sequences and character state changes.
- Compare tree lengths.
- Using **hiv-phy.txt**, and **DNAML** program, obtain ML tree with F84 distances.
- Select the correct options to estimate ancestral sequences.
- Compare likelihoods values.

4. **Phylogenies using MEGA**.

   - Explore MEGA3.0 facilities using Drosophila ADH example.
   - See *Data explorer* and *Statistics*
   - Compute LS, ME, MP and NJ trees.

5. **Likelihood using TREE-PUZZLE**.

   - Using **hiv1-phy.txt** and **mtv1-phy.txt** obtain ML tree under HKY+Γ model using 8 rate categories.
   - Observe ML distance matrix. Sequence composition test. Ts:Tv ratio estimation. Observe Likelihood value. α estimation.
   - Using a treefile with 4 alternative topologies (**intree.txt**) compute **KHT**, **SHT** and **ELW** test.

- Make an intree file for **hiv1-phy.txt** sequences and compute the above paired site tests.

6. **MODELTEST**.

   - Take your time to see the **mtv1-mdt.nex** file format.
   - Run **mtv1-mdt.nex** using **PAUP***[26].
   - Run **MODELTEST** using **model.score** file.
   - bin > Modeltest3.5.win -d2 < mtv1-model.score > mtv1-model.out
   - What is the best model of evolution for the data set?

7. **Bayesian using Mr Bayes**.

   - Use the **hiv1-by.txt** file format.
   - Take your time to see the file format.
   - Run MrBayes typing **execute hiv1-by.txt**
   - Compare parameters estmated by MrBayes and Modeltest

[26] Since PAUP* is not free (although not expensive) an alternative is to use **PAML** package.

13. 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/
  - **Phylemon at CIPF** http://bioinfo.cipf.es/cgi-bin/mortadelo/cgi/tools.cgi

14. Credits

This presentation is based on:[27]

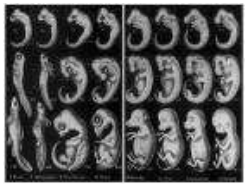
- Major Book or Chapters References:
  - Swofford, D. L. *et al.* **1996**. Phylogenetic inference [101].
  - Harvey, P. H. *et al.* **1996**. New Uses for New Phylogenies [36].
  - Page, R. & Holmes, E. **1998**. Molecular evolution. A phylogenetic approach [36].
  - Li, W. S. **1997** . Molecular Evolution [60].
  - Hartl, D. & Clark, A. **1999** . Principles of population genetics [35].
  - Nei, M. & Kumar, S. **1999** . Molecular evolution and phylogenetics [74].
  - Salemi, M. & Vandamme, A. (ed.) **2003**. The phylogenetic handbook [89].
  - Balding, Bishop & Cannings. (ed.) **2003**. Handbook of Statistical Genetics [2].
  - Felsenstein, J. **2004**. Inferring phylogenies [22].
  - Nielsen, R. (ed.) **2004**. Statistical Methods in Molecular Evolution [15].
- 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://filogeografia.dna.ac/PDFs/phylo/Foster_01_EasyIntro_MLPhylo.pdf

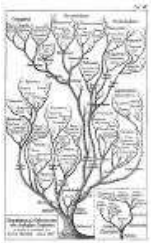
[27] Latex and pdfscreen package. HJD take responsibility for innacuracies of this presentation.

15. Additional Material

15.1. What are the roots of modern phylogenetics?

Phylogenies have been inferred by systematics since Darwin and Haeckel,



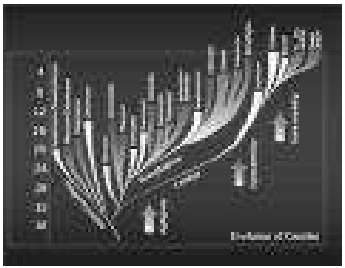


However, since 1950s-60s classifications began to be more numerical, algorithmic and statistical. Principally due to progress in molecular biology, protein sequence

data and computer development (initially, using punched card machines) [28].
Roughly, systematists divided in two:

1. Proponents of the **"Evolutionary Systematics"** classify organisms using different historical, ecological, numerical, and evolutionary arguments. It attemps to represent, not only the branching of phyletic lines (cladogenesis) but also its subsequent divergence (anagenesis) leading the invasion of a new adaptive zone by a particular class of organisms (a grade). Its representaties are Ernst Mayr[65] and George G. Simpson[94], among others.

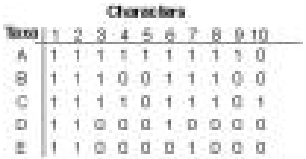




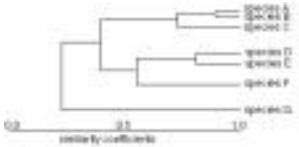
2. Proponents who rejected the notion of theory-free method of classification, introduced **objectivity** by using explicit numerical approaches.

[28] See: Chapter 5 of [66] and Chapter 10 of [22] for a detailed discussion on the issue.

(a) Numerical Taxonomy's school (**Phenetics**) originated by Michener[68], Sneath[96] and Sokal[97] in USA.

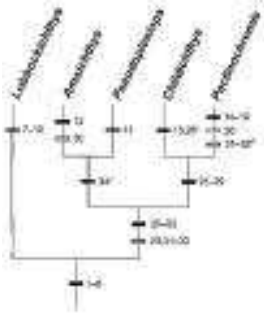


Taxa	Characters									
	1	2	3	4	5	6	7	8	9	10
A	1	1	1	1	1	1	1	1	1	0
B	1	1	1	0	0	1	1	1	0	0
C	1	1	1	1	0	1	1	1	0	1
D	1	1	0	0	0	1	0	0	0	0
E	1	1	0	0	0	0	1	0	0	0



- **Main idea:**
  To score pairwise differences between OTU's (Operational Taxonomic Units) using as many characters as possible.
  Cluster by simmilarity using an algorithm that produces a single dendogram (**phenogram**)

(b) Phylogenetic Systematic's school (**Cladistics**) originated by Hennig[42, 43] in Germany and followed by Wagner[103], Kluge[54] and Farris[17, 18] in USA.

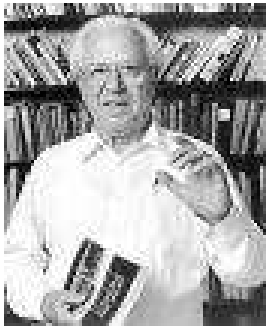


- **Main idea:**
  To use recency of common ancestry to construct hierarchies of relationship, NOT similarity.
  Relationships depicted by phylogenetic tree, show sequence of speciation events (**cladogram**)[29].

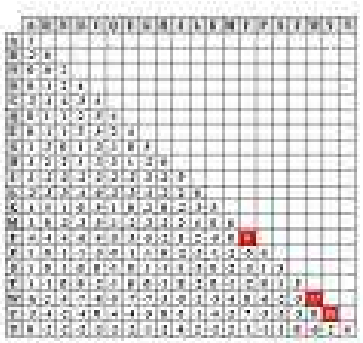
[29] Felsenstein[22] asserts that although Edwards and Cavalli-Sforza introduced parsimony, modern work on it springs from the paper of Camin and Sokal[8]

(c) Statistical approaches developed around molecular data sets.

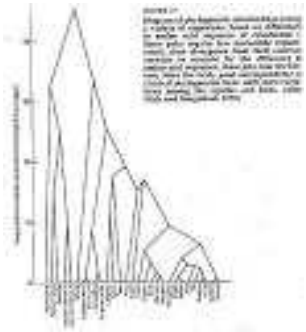
- Edwards and Cavalli-Sforza[9, 10] worked on the spatial representation of human gene frequencies differences, developed the **Minimum Evolution** and the **Least Square** distance methods, respectively. In order to reconcile results, they worked out an impractical **Maximum Likelihood** method and found that it was not equivalent to either of their two methods! Indeed, they discussed similarities between a **Maximum Parsimony** method and likelihood [9].



- In the 1960s the molecular sequence data was mostly proteins. Margareth Dayhoff began to accumulate in **the first molecular database!** produced in a printed form [14]. In the second edition of the "Atlas..." they describe the **first molecular parsimony method**, based on a model in wich each of the 20 amino acids was allowed to change to any of the 19 others in a single step **(unordered method)**.

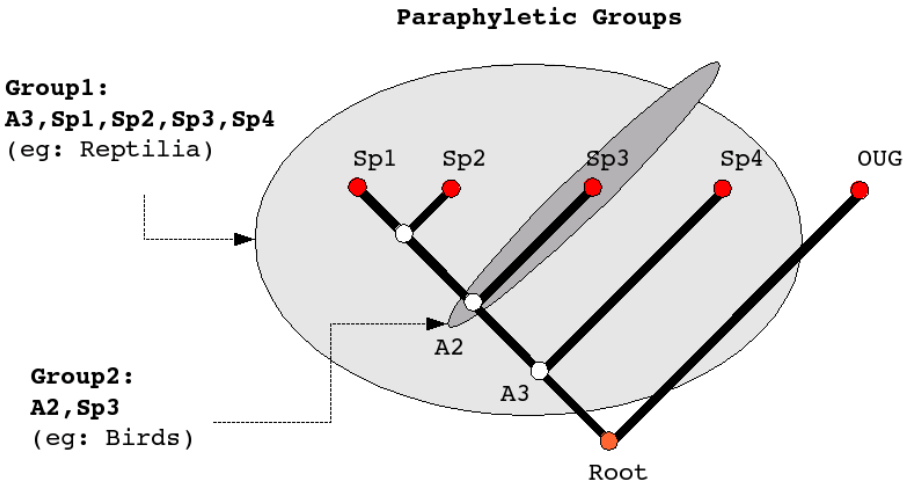


- Although distance methods were first described by Edwards and Cavalli-Sforza [9, 10], Fitch and Margoliash [28] popularized distance matrix methods based on **least squares**. The distances were fractions of amino acids differences between a particular pair of sequences. The least squares was weighted with greather observed distance given less weight. **This introduces the concept that large distances would be more prone to random error owing to the stochasticity of evolution.**



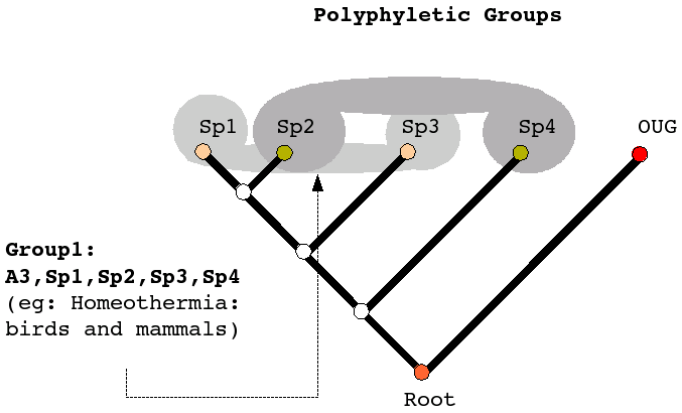
- Explicit models of sequence evolution correcting the effects of **multiple replacement** was first implemented by Jukes and Cantor in 1969 [50].

Paraphyletic group represents a group of organisms derived from a single ancestral taxon, but one which does not contain all the descendants of the most recent common ancestor[30].



[30]Paraphyly derives from the evolutionary differentiation of some lineages, based on the accumulation of specific autapomorphies (eg: Birds)

Polyphyletic group represents a group of organisms with the same taxonomic title derived from two or more distinct ancestral taxa[31]. Frequently, paraphyletic or polyphyletic groups are considered **grades**[32]



Sometimes is difficult to distinguish clearly between artificial groups.

The important contrast is between monophyletic and nonmonophyletic groups!!

[31] Polyphyly derives from convergence, paralelisms or reversion (homoplasy) rather than common ancestry (homology)

[32] It is an evolutionary concept supposed to represent a taxon with some level of evolutionary progress, level of organization or level of adaptation

15.2. Types of data

All of the experimental data gathered by molecular biologists fall into one of the two broad categories: **discrete characters** and similarities or **distances**.

- A discrete character provides data about an individual species or sequences.
- Character data are often transformed into distances.
- Discrete character data are those for which a data matrix X assigns a **character state** x_{ij} to each taxon i for each character j.
- Characters may be binary or multistate.
- Multistate characters may be ordered or unordered, depending on whether an ordering relationship is imposed upon the possible states

- The concepts of **character order** and **character polarity** should not be confused. The former defines the allowed character-states transformations, whereas the later refers to the **direction** of evolution.
- Nucleotide sequence data are generally treated as unordered multistate characters, since there is no *a priori* reasons to assume, for example, that state C is intermediate between A and G.

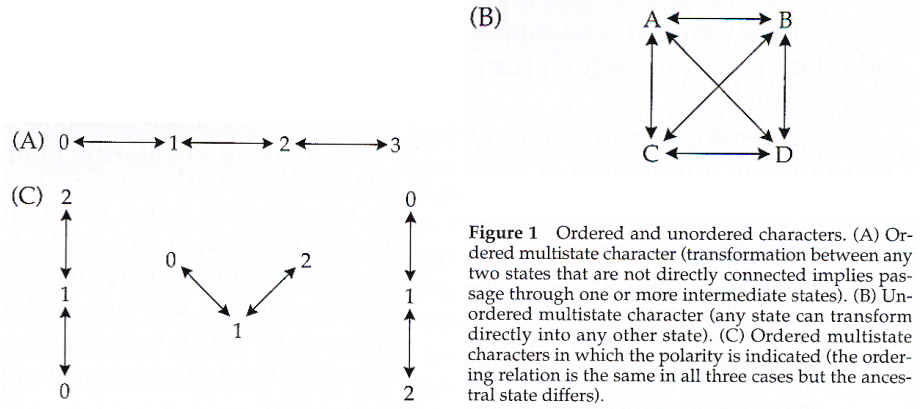


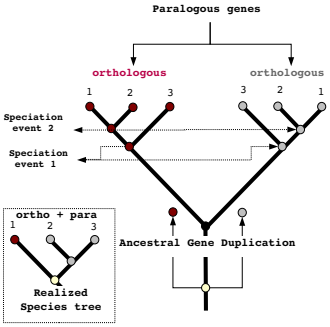
Figure 1 Ordered and unordered characters. (A) Ordered multistate character (transformation between any two states that are not directly connected implies passage through one or more intermediate states). (B) Unordered multistate character (any state can transform directly into any other state). (C) Ordered multistate characters in which the polarity is indicated (the ordering relation is the same in all three cases but the ancestral state differs).

15.3. 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[33] or true **species tree**.

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



2. In presence of **polymorphic alleles** at a locus, the time of gene splitting (producing polimorphisms) is usually earlier than population or species splitting.

The probability to obtain the expected species tree depends on T & N and random processes like lineage sorting [77].

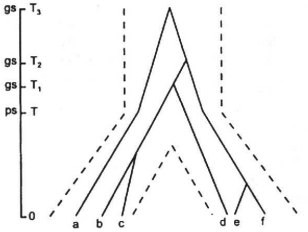


FIGURE 5.2. Diagram showing that the time of gene splitting (gs) is usually earlier than the time of population splitting (ps) when polymorphism exists. From Takahata and Nei (1985).

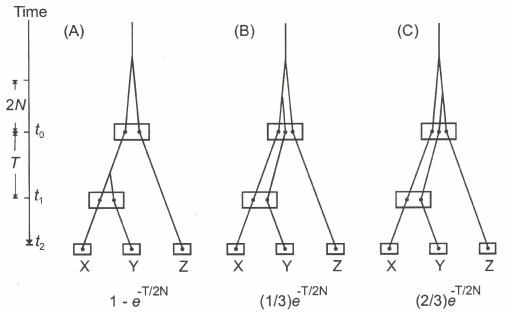
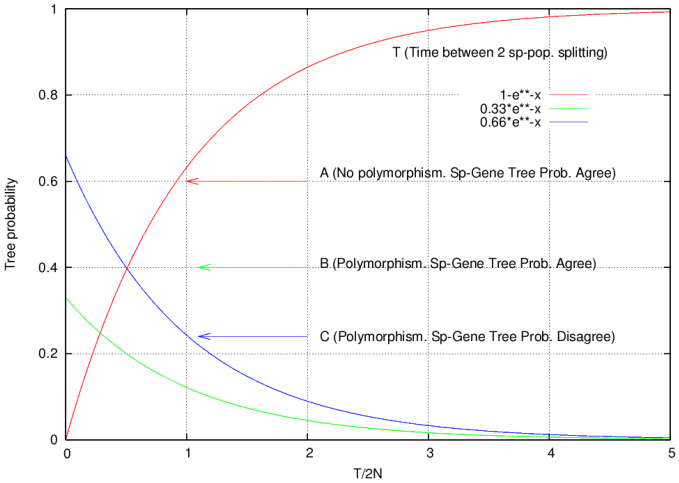
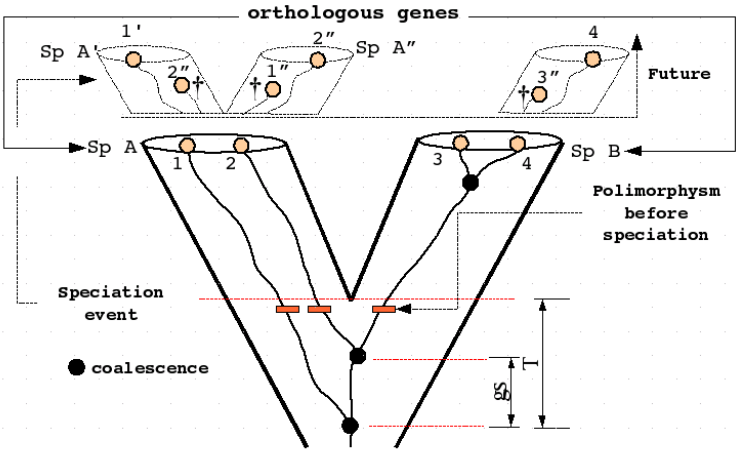


FIGURE 5.3. Three possible relationships between the species and gene trees for the case of three species in the presence of polymorphism. The times of the first and second species splitting are t_0 and t_1, respectively. The probability of occurrence of each tree is given underneath the tree. $T = t_1 - t_0$, and N is the effective population size. From Nei (1987).

- If alleles are monophyletic before population or species splitting, at time T/2N increase (longer times or low pop. numbers-*mammals*-), the probability to agree between trees increases (red, A tree pattern).
- This probability decreases if polymorphic alleles are present before the pop. splitting. For a constant T value, increasing population size reduces the probability of random processes reducing polymorphism (green, B tree pattern).
- In such conditions the probability of disagreement between trees is higher (blue, C tree pattern).



- Indeed future sorting events could prevent the correct tree gene.



Alleles must be monophyletic at time of speciation to obtain a reliable species tree

Sometimes there are local clocks

for example mouse and rat using (*hamster as outgroup*)[34]

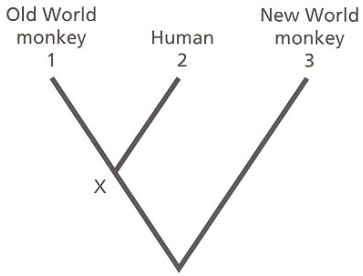
[34] See [4] for an actualized review.

TABLE 8.1 Numbers of nucleotide substitutions per 100 sites between species[a]

	Synonymous sites		Nonsynonymous sites	
Species pair	K_S	L^b	K_A	L^b
Mouse–rat	18.0 ± 0.7	4,229	1.8 ± 0.1	15,217
Mouse–hamster	30.3 ± 1.0	4,229	2.9 ± 0.1	15,217
Rat–hamster	31.3 ± 1.0	4,229	2.7 ± 0.1	15,217
Mouse–human	53.4 ± 1.5	4,229	5.2 ± 0.2	15,217
Rat–human	51.6 ± 1.5	4,229	5.0 ± 0.2	15,217
Hamster–human	52.3 ± 1.5	4,229	5.1 ± 0.1	15,217

Relative Rate Test

How to test the molecular clock?[35]



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)
$d_{12} = 6.9$
$d_{13} - d_{23} = 1.0 \pm 0.5$

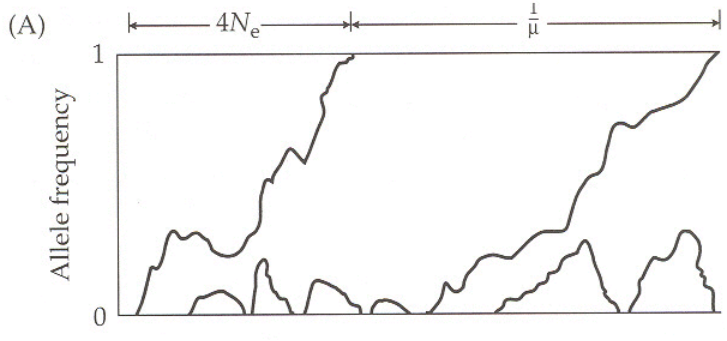
(d) Two flanking regions (936 bp)
$d_{12} = 7.9$
$d_{13} - d_{23} = 3.1 \pm 1.1$*

[35] See [84] and download RRtree!!

15.4. Neutral theory of evolution

At molecular level, the most frequent changes are those involving fixation in populations of neutral selective variants [52].

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



15.5. Ultrametric & Additive Properties

Distance to be represented in a tree diagram must be **metric** and **additive**. Let $d(a, b)$ the distance between 2 sequences, d is metric if:

1. $d(a, b) \geq 0 \mapsto$ (non-negative),
2. $d(a, b) = d(b, a) \mapsto$ (symmetry),
3. $d(a, c) \leq d(a, b) + d(b, c) \mapsto$ (**triangle inequality**),
4. $d(a, c) = 0$ if and only if $a = b \mapsto$ (distinctness)

   ♣ A metric is an ultrametric if it satisfies the additional criterion that:

5. $d(a, b) \geq maximum[d(a, c), d(b, c)] \mapsto$ (the two largest distance are equal),

   ♣ Being metric (or ultrametric) is a necessary but not sufficient condition for being a valid measure of evolutionary change. A measure must also satisfy the **the four-point condition**:

6. $d(a, b) + d(c, d) \leq maximum[d(a, c) + d(b, d), d(a, d) + d(b, c)]$

15.6. Optimality Criteria

Inferring a phylogeny is an estimate procedure.

We are making **a "best estimate"** of an evolutionary history based **on the incomplete information** contained in the data.

Because we can postulate evolutionary scenarios by which any chosen phylogeny could have produced the observed data, **we must have some basis for selecting one or more preferred trees** among the set of possible phylogenies.

As we have seen, we can define a specific algorithm that leads to the determination of a tree, but also, we can define **a criterion for comparing alternative phylogenies to one another and decide which is better**.

Cluster analysis methods combine tree inference and the definition of the preferred tree **into a single** statement. In fact, UPGMA and NJ give a single tree.

Methods using optimality criterion has **two logical steps**.

The **first** is to define an objetive function to **score trees**, and the **second** is to **find alternative trees** to apply the criterion. The last problem will be covered below the title: **"searching trees"**.

This kind of procedure would produce **many alternative optimal solu-**

tion.

15.6.1. Least squares family methods

We can now address the problem of choosing a tree from the following conceptual perspective: *We have uncertain data that we want to fit to a particular mathematical model* **(and additive tree)** *and find the optimal value for the adjustable parameters* **(the topology and the branch lengths)**.

Several methods depend on a definition of the disagreement between a tree and the data based on the following familiy of objective functions:

$$E = \sum_{i=1}^{T-1} \sum_{j=i+1}^{T} w_{ij} \mid d_{ij} - p_{ij} \mid^{\alpha}$$

Where E defines the error of fitting the distance estimates to the tree, T is the number of taxa, w_{ij} is the weight applied to the separation of taxa i and j, d_{ij} is the pairwise distance estimate (*matrix distances*), p_{ij} is the length of the path connecting i and j in the given tree[36], the vertical bars represent absolute values, and $\alpha = 1$ or 2.

Methods depend on the selection of specific α and the weighted scheme w_{ij}

[36] p_{ij} is also called as **patristic distances**

- If $\alpha = 2$ and $w_{ij} = 1$, the unweighted squared deviations will be minimized, assuming that all the distance estimates are subject to the same magnitude of error (LS of C-S&E)[10].
- If $\alpha = 2$ and $w_{ij} = 1/d_{ij}^2$, the weighted squared deviations will be minimized, assuming that the estimates are uncertain by the same percentage (LS method of F&M)[28].

15.6.2. Minimum Evolution

The minimum evolution method [51, 86, 87, 88] uses a criterion:

the total branch length of the reconstructed tree.

$$S = \sum_{k=1}^{2T-3} \mid v_k \mid$$

That is, the optimality criterion is simply the sum of the branch lengths that minimize the sum of squared deviations between the observed (estimated) and path-length (patristic) distances.

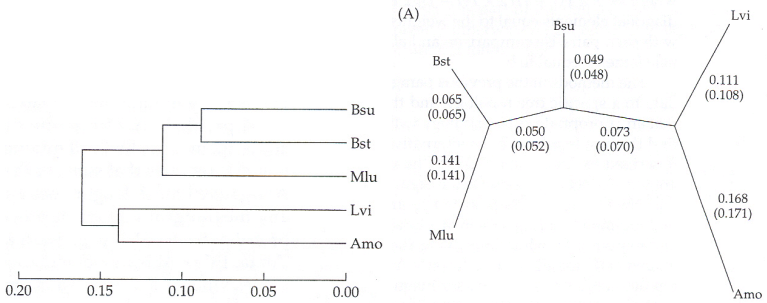
Thus this method makes partial use of the LS (C-S&E) criterion.

Under the ME criterion, a tree is worse than another tree only if its S value is **significantly larger** than that of the other tree.

Thus, all trees whose S values are not significantly different from the minimum S value should be regarded as candidates for the true tree[37].

Rzhetsky & Nei [86] proposed a fast approximated search of the ME tree based on the observation that ME tree (*below*) is almost always identical to NJ tree.

UPGMA NJ & (LS) methods and values of expected substitutions per sequence position



[37]The statistical procedure for testing different trees will be discussed in "confidence on trees".

15.7. Parsimony Criteria

A common misconception regarding the use of parsimony methods is that they require *a priori* determination of character polarities.

In morphological studies, character polarity is commonly inferred using **outgroup comparison**, however, it is by no means a prerequisite to the use of parsimony methods.

Parsimony analysis actually compromises a group of related methods differing in their underlying evolutionary assumptions.

- **Wagner Parsimony** [54, 18] ordered, multistate characters with reversibility.
- **Fitch Parsimony** [25] unordered, multistate characters with reversibility.

	a	b	c	d
a	–	1	2	3
b	1	–	1	2
c	2	1	–	1
d	3	2	1	–

	a	b	c	d
a	–	1	1	1
b	1	–	1	1
c	1	1	–	1
d	1	1	1	–

- Since both Fitch and Wagner Parsimony allow reversibility, the tree may be rooted at any point without changing the tree length.

- **Dollo Parsimony** [13], reversals allowed, but the derived state may arise only once [38]

	a	b	c	d
a	–	M	$2M$	$3M$
b	1	–	M	$2M$
c	2	1	–	M
d	3	2	1	–

- **Transversion Parsimony** [6], transition substitutions (Pu↦Pu; Py↦Py) occur more frequently than transversion (Pu↦Py; Py↦Pu) substitutions. Pu(A,G); Py(C,T).

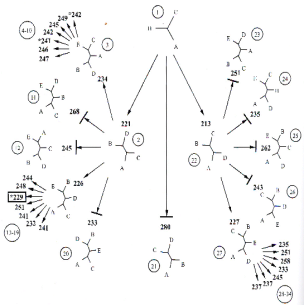
	A	C	G	T
A	–	5	1	5
C	5	–	5	1
G	1	5	–	5
T	5	1	5	–

[38] Dollo Parsimony is suggested for restriction site data or for very complex characters that probably have only arisen once, such as legs in tetrapods or wings in insects. M is an arbitrary large number, guaranteeing that only one transformation to each derived state will be permitted.

15.8. Searching Trees.

Branch & Bound search[40]

- Much faster, but still **guaranteed to find the best tree**,
- Determine an **upper bound for the shortest tree**,
  - Use the length of **a random tree**
- Follow a predictable search path through possible tree topologies, **similar to an exhaustive search**,
- **Abandon** any fork of the search tree **when the upper bound is exceeded before the last taxon is added**,
- **Does not calculate the length of all trees but finds the best one**



Star Decomposition

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

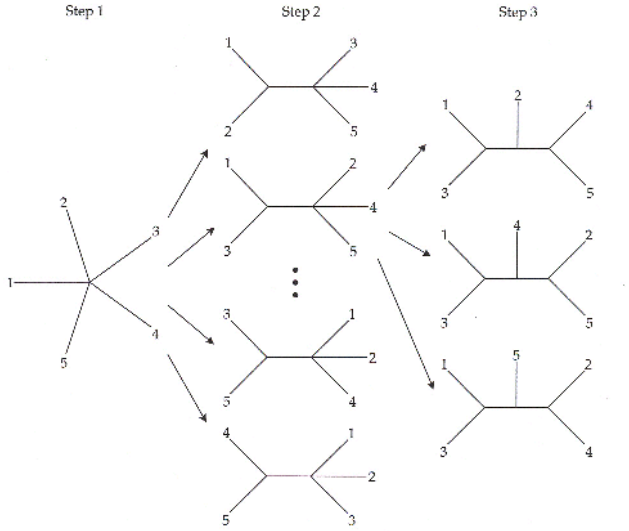


Figure 25 Heuristic tree selection using star decomposition method. At each step, the optimality criterion is evaluated for each possible joining of a pair of lineages leading away from the central node. The best tree found during each step becomes the starting point for the next step.

Nearest Neighbor Interchange

- Identify an interior branch. It is flanked by four subtrees
- Swap two of the subtrees on opposite ends of the branch
- Two rearrangements are possible

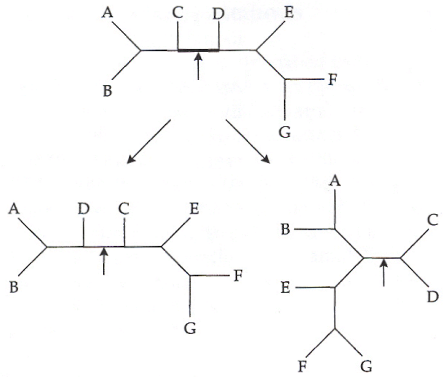


Figure 26 Branch swapping by nearest-neighbor interchanges (NNIs). Each interior branch of the tree defines a local region of four subtrees connected by the interior branch. Interchanging a subtree on one side of the branch with one from the other constitutes an NNI. Two such rearrangements are possible for each interior branch.

Subtree Pruning & Regrafting

- Identify and remove a subtree
- Reattach to each possible branch of the remaining tree
- NNI is a subset of SPR

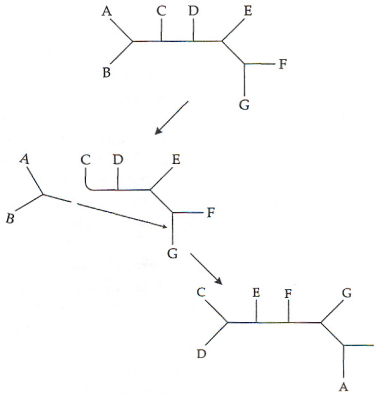


Figure 27 Branch swapping by subtree pruning ar regrafting. A subtree is pruned from the tree (e.g., tl subtree containing terminal nodes A and B as ind cated). The subtree is then regrafted to a different loc tion on the tree. All possible subtree removals and rea tachment points are evaluated.

15.9. Molecular adaptation

A powerfull approach to detecting molecular evolution by positive (Darwinian) selection derives from comparison of the relative rates of synonymous and non-synonymous substitutions (citar)[39].
Synonymous mutations do not change the amino acid sequence; hence their substitution rates (dS) is **"neutral"**[40] with respect to selective pressure on the protein product.
Nonsynonymous mutations do change the amino acid sequence, so their substitution rate (dN) is a function of selective pressure on the protein.
The ratio of these rates ($\omega = dN/\ dS$) is a function of selective pressure.

If nonsynonymous mutations are deleterious, **purifying selection** will reduce their fixation rate and $dN/dS < 1$.

If nonsynonymous mutations are advantageous **adaptive**, they will be fixed at a higher rate than synonymous mutations, and $dN/dS > 1$.

A $dN/dS = 1$ is consistent with **neutral evolution**.

[39] This section is largely based on [109]
[40] See [11] for a discussion about this issue

15.9.1. Counting methods

We wish to estimate the number of synonymous substitutions per synonymous site (dS) and the number of nonsynonymous substitutions per nonsynonymous site (dN) between two protein-coding sequences. In the past two decades, about a dozen methods have been proposed for this estimation. They are intuitive and involve treatment of the data that cannot be justifief rigoroursly.

All counting methods roughly work like this:
Suppose the gene has **300 codons** and we observe **5 synonymous and 5 nonsynonymous** differences.
Can we conclude that synonymous and nonsynonymous substitution rates are equal with $\omega = 1$?...NO!

An inspection of the genetic code table suggests that **all changes in the second position and most changes at the first are nonsynonymous**, and **only some changes at the third position are synonymous**. Consequently we do not expect synonymous and nonsynonymous mutations at equal proportions even if there is no selection at the protein level.

Indeed, if mutations from any one nucleotide to any other occur at the same rate, we expect 25.5% of mutations to be synonymous and 74.% to be nonsynonymous [112].

If we use those proportions, it is clear that selection on proteins has decreased the fixation rate of nonsynonymous mutations by about 3 times, since $\omega = 5/5/(74.5/25.5) = 0.34$

There are 900 nucleotides in the gene, so the number of synonymous (S) and nonsynonymous (N) sites are S=900 x 25.5%=229.5 and N=900 x 74.5%=670.5, respectively. Then, we have dS=5/229.5=0.0218 and dN=5/670.5=0.0075.

Therefore counting methods involve 3 steps:

- 1. Count the number of **sites** S and N in the two cDNA sequences
  - Complicated by factors such as ts/tv rate bias and base /codon frequency bias.
- 2. Count the number of synonymous and nonsynonymous **differences**
  - This is straigthtforward if the two compared codons differ at one codon position only. When they differ at 2 or 3 codon positions , there exists 4 or 6 pathways from one codon to the other. The multiple pathways may involve different number of synonymous and nonsynonymous and should ideally be weighted appropriately according to their likelihood of occurrence. Most counting methods use equal weighting
- 3. Apply a correction for multiple substitution at the same site.

– Counting methods use multiple-hit correction formulas based on nucleotide -substitution models, assuming nucleotides change to 1 of 3 other nucleotides. When those formulas are applied to synonymous (or nonsynonymous) sites only.

The method of Miyana-Yasunaga [69] and its simplified version (Nei-Gojobori [73]) are based on nucleotide substitution model of Jukes and Cantor [50]) and ignore the ts/tv bias or base codon frequency.
Since ts are more likely to be synonymous than tv at 3rd. position, ignoring the ts/tv rate bias understimate the number of S and overestimate N. This effect is well known, and different methods account for this ratio (Li et al. [59], Li [58], Pamilo and Bianchi [78], Ina [48].)

The effect of biased base/codon frequencies can have devastating effects on the estimation of dN and dS . Qualitatively different conclusions were reached dpending on wether codon usage bias is accomodated for nucler genes from mammals and Drosophila [3].

A counting method incorporating both the ts/tv bias and the base/codon frequency bias was implemented by Yang and Nielsen [110].
Many, if not all of them, are incorporated in codeml(PAML) [108].

15.9.2. Markov model of codon substitutions

In molecular phylogenetics we use a Markov process to describe the change between nucleotides, amino acids, or codons over evolutionary time [61, 72].

Perviously we describe evolutionary models based on different Markovian processes (DNA or amino acid models). Now we describe **substitutions between the sense codons**. Stop codons are excluded. The "Universal" genetic code, there are **61 sense codons** (and 3 stops), therefore 61 states in the Markov process.

The Markov process is characterized by a rate matrix $Q = \{q_{ij}\}$, where q_{ij} is the substitution rate from sense codon i to sense codon j $(i \neq j)$. Formally, $q_{ij}\Delta t$ is the probability that the process is in state j after an infinitesinal time Δt, given that it is in state i at time t.

The model accounts for **ts/tv bias, unequal synonymous and nonsynonymous substitution, and biased base/codon frequencies**. Mutations are assumed to occur independently among the 3 codon positions, and so only one position is allowed to change instantaneously. Since ts occur more frequently than tv, the model multiply the rate by ts/tv rate ratio κ if the chage is a transition. To account for codon usage bias, the model let π_j be the equilibrium frequency of codon j and multiply substitution rates to codon j by π_j.

The model can either use all π_j as parameters, with 60 (61-1) free parameters used, or calculate π_j from base frequency at the 3 coson positions, with 9=3x(4-1) free parameters used.

To account for synonymous and nonsynonymous substitution rates, the model multiply the rate by ω if the change is nonsynonymous. It is important to note that that parameters κ and π_j characterize processes, including selection, **at the DNA level**, while selectionat **the protein level** has the effect of modifying parameter ω. If natural selection operates on the DNA as well as on the protein, the synonymous rate will differ from the mutation rate.

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at 2 or 3 codon position,} \\ \mu\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous tv,} \\ \mu\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous ts,} \\ \mu\omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous tv,} \\ \mu\omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous ts,} \end{cases}$$

For example, consider the substitution rates **to** codon CTG (Leu). We have
$q_{CTC,CTG} = \mu\pi_{CTG}$ since the CTC(Leu)→ CTG(Leu) change is a syn tv,
$q_{TTG,CTG} = \mu\kappa\pi_{CTG}$ since the TTG(Leu)→ CTG(Leu) change is a syn ts,
$q_{GTG,CTG} = \mu\omega\pi_{CTG}$ since the GTG(Val)→ CTG(Leu) change is a nonsyn tv,
$q_{CCG,CTG} = \mu\kappa\omega\pi_{CCG}$ since the CCG(Pro)→ CTG(Leu) change is a nonsyn ts
$q_{TTT,CTG} = 0$ since the TTT and CTG differ at 2 positions

The diagonal elements of the matrix $Q = \{q_{ij}\}$ are determined by mathematical requirements that each row in the matrix sums to zero.

$$\sum_j q_{ij} = 0, \text{for any } i$$

Molecular sequence data do not allow separate estimation of the rate (μ) and time (t), and only their product (μt) can be identified. We thus fix the rate μ such that the expected number of nucleotide substitutions per codon is one:

$$-\sum_i \pi_i q_{ii} = \sum_i \pi_i \sum_{j \neq i} q_{ij} = 1$$

This scaleng means that time t is measured by distance, the expected number of (nucleotide) substitutions per codon. The transition probability matrix over time t is

$$P(t) = \{p_{ij}(t)\} = e^{Qt},$$

Lastly, the model is time - reversible. This means,

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \text{ for any } t, i \text{ and } j$$

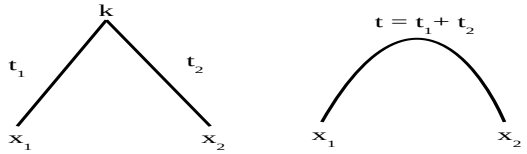
15.9.3. Maximum likelihood estimation

Below we[41] describe the ML method for estimating dN and dS (Goldman and Yang[32]). The data are two aligned protein-coding DNA sequence,

```
Human   GAG CCC TGG CCT CTC ...
Mouse   GAG CTC TCG ACT GTT ...
```

We assume that all the codons are evolving independently according to the same Markov process. Suppose there are n **sites (codons)** in the gene, and let the **data at site** h **be** $x_h = \{x_1, x_2\}$, where x_1 and x_2 are the two codons in the sequences at that site.
In the example, the data at site $h = 2$ are x_1 =CCC, x_2 =CTC. **The probability of observing data** x_h **at site** h is,

$$f(x_h) = \sum_{k=1}^{61} \pi_k p_{kx_1}(t_1) p_{kx_2}(t_2)$$



Parameter t_1 and t_2 cannot be estimated separately, only their sum is estimable.

[41] Remember we are following Yang[109]

$$f(x_h) = \sum_{k=1}^{61} \pi_k p_{kx_1}(t_1) p_{kx_2}(t_2) = \pi_{x_1} p_{x_1 x_2}(t_1 + t_2)$$

Parameters in the model are: the sequence divergence t, the transition/transversion rate ratio κ, the nonsynonymous/synonymous rate ratio ω, and the codon frequency π_j. The log-likelihood function is then given by

$$l(t, \kappa, \omega) = \sum_{n=1}^{n} log f(x_h)$$

Codon frequencies ($\pi_i' s$) can usually be estimated by using observed base or codon frequencies. Since **there is not an analitical solution**, a numerical hill-climbing algorithm is used to maximize the l

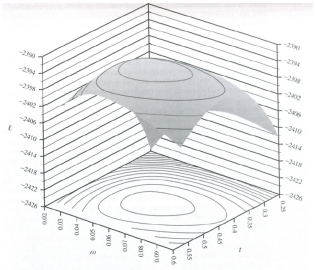


Figure 9.2 The log-likelihood surface contour as a function of parameters t and ω for the comparison of the human and mouse acetylcholine receptor α genes. The maximum likelihood method estimates parameters by maximizing the likelihood function. For these data, the estimates are $\hat{t} = 0.444$, $\hat{\omega} = 0.059$, with optimum log-likelihood $\hat{\ell} = -2392.83$.

The table shows the estimations of different counting methods and ML estimation for a **pairwise comparison of sequences**.

Table 5.1. Estimation of d_S and d_N between *Drosophila melanogaster* and *D. simulans GstD1* genes.

Method	κ	S	N	d_S	d_N	ω	ℓ
ML methods							
Fequal, $\kappa = 1$	1	152.9	447.1	0.0776	0.0213	0.274	-927.18
Fequal, κ estimated	1.88	165.8	434.2	0.0221	0.0691	0.320	-926.28
F3×4, $\kappa = 1$	1	70.6	529.4	0.1605	0.0189	0.118	-844.51
F3×4, κ estimated	2.71	73.4	526.6	0.1526	0.0193	0.127	-842.21
F61, $\kappa = 1$	1	40.5	559.5	0.3198	0.0201	0.063	-758.55
F61, κ estimated	2.53	45.2	554.8	0.3041	0.0204	0.067	-756.57
Counting methods							
Nei and Gojobori	1	141.6	458.4	0.0750	0.0220	0.288	
Yang and Nielsen (F3×4)	3.28	76.6	523.5	0.1499	0.0190	0.127	

15.9.4. Phylogenetic estimationm of selective pressure

Table 9.2 Log-likelihood values and parameter estimates under different models for the lysozyme *c* genes.

Model	p	ℓ	$\hat{\kappa}$	$\hat{\omega}_0$	$\hat{\omega}_h$	$\hat{\omega}_c$
A. 1 ratio: $\omega_0 = \omega_h = \omega_c$	22	−906.02	4.5	0.81	$= \hat{\omega}_0$	$= \hat{\omega}_0$
B. 2 ratios: $\omega_0 = \omega_h, \omega_c$	23	−904.64	4.6	0.69	$= \hat{\omega}_0$	3.51
C. 2 ratios: $\omega_0 = \omega_c, \omega_h$	23	−903.08	4.6	0.68	∞	$= \hat{\omega}_0$
D. 2 ratios: $\omega_0, \omega_h = \omega_c$	23	−901.63	4.6	0.54	7.26	$= \hat{\omega}_H$
E. 3 ratios: $\omega_0, \omega_h, \omega_c$	24	−901.10	4.6	0.54	∞	3.65
F. 2 ratios: $\omega_0 = \omega_h, \omega_c = 1$	22	−905.48	4.4	0.69	$= \hat{\omega}_0$	1
G. 2 ratios: $\omega_0 = \omega_c, \omega_h = 1$	22	−905.38	4.4	0.68	1	$= \hat{\omega}_0$
H. 2 ratios: $\omega_0, \omega_h = \omega_c = 1$	22	−904.36	4.3	0.54	1	1
I. 3 ratios: $\omega_0, \omega_h, \omega_c = 1$	23	−902.02	4.5	0.54	∞	1
J. 3 ratios: $\omega_0, \omega_h = 1, \omega_c$	23	−903.48	4.4	0.54	1	3.56

Note: p is the number of parameters. All models include the following 21 common parameters: 11 branch lengths in the tree (Figure 9.5), 9 parameters for base frequencies at codon positions used to calculate codon frequencies, and the transition/transversion rate ratio κ. *Source*: Yang (1998).

Table 9.3 Likelihood ratio statistics ($2\Delta\ell$) for testing hypotheses concerning lysozyme evolution.

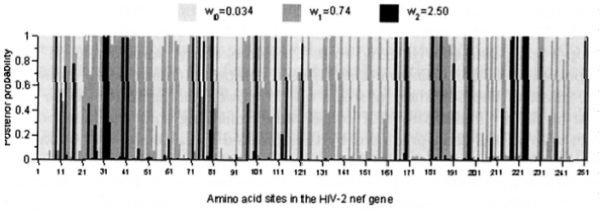
Hypothesis tested	Assumption made	Models compared	$2\Delta\ell$
A. $(\omega_h = \omega_c) = \omega_0$	$\omega_h = \omega_c$	A & D	8.78**
B. $\omega_c = \omega_0$	$\omega_h = \omega_0$	A & B	2.76
C. $\omega_c = \omega_0$	ω_h free	C & E	3.96*
D. $\omega_h = \omega_0$	$\omega_c = \omega_0$	A & C	5.88*
E. $\omega_h = \omega_0$	ω_c free	B & E	7.08**
A′. $(\omega_h = \omega_c) \leq 1$	$\omega_h = \omega_c$	D & H	5.46*
B′. $\omega_c \leq 1$	$\omega_h = \omega_0$	B & F	1.68
C′. $\omega_c \leq 1$	ω_h free	E & I	1.84
D′. $\omega_h \leq 1$	$\omega_c = \omega_0$	C & G	4.60*
E′. $\omega_h \leq 1$	ω_c free	E & J	4.76*

*Significant at the 5 % level ($\chi^2(1) = 3.84$).
**Significant at the 1 % level ($\chi^2(1) = 6.63$).
Source: Yang (1998).

15.9.5. Adaptive evolution on amino acid sites

Table 5.3. Parameter estimates and likelihood scores under models of variable ω ratios among sites for HIV-2 *nef* genes. (Note: The number after the model code, in parentheses, is the number of free parameters in the ω distribution. The d_N/d_S ratio is an average over all sites in the HIV-2 *nef* gene alignment. Parameters in parentheses are not free parameters and are presented for clarity. PSS is the number of positive selected sites, inferred at the 50% (95%) posterior probability cutoff.)

Model	d_N/d_S	Parameter estimates	PSS	ℓ
M0: one ratio (1)	0.51	$\omega = 0.505$	none	−9775.77
M3: discrete (5)	0.63	$p_0 = 0.48, p_1 = 0.39, (p_2 = 0.13)$ $\omega_0 = 0.03, \omega_1 = 0.74, \omega_2 = 2.50$	31 (24)	−9232.18
M1: neutral (1)	0.63	$p_0 = 0.37, (p_1 = 0.63)$ $(\omega_0 = 0), (\omega_1 = 1)$	not allowed	−9428.75
M2: selection (3)	0.93	$p_0 = 0.37, p_1 = 0.51, (p_2 = 0.12)$ $(\omega_0 = 0), (\omega_1 = 1), \omega_2 = 3.48$	30 (22)	−9392.96
M1a: nearly neutral (2)	0.48	$p_0 = 0.55, (p_1 = 0.45)$ $(\omega_0 = 0.06), (\omega_1 = 1)$	not allowed	−9315.53
M2a: positive selection (4)	0.73	$p_0 = 0.51, p_1 = 0.38, (p_2 = 0.11)$ $(\omega_0 = 0.05), (\omega_1 = 1), \omega_2 = 3.00$	26 (15)	−9241.33
M7: beta (2)	0.42	$p = 0.18, q = 0.25$	not allowed	−9292.53
M8: beta & ω (4)	0.62	$p_0 = 0.89, (p_1 = 0.11)$ $p = 0.20, q = 0.33, \omega = 2.62$	27 (15)	−9224.31



References

[1] J. Adachi and M. Hasegawa. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J Mol Evol*, 42:459–468, 1996.

[2] D. Balding, M. Bishop, and C. Cannings (eds.). *Handbook of Statistical Genetics*. Wiley J. and Sons Ltd., N.Y., second edition, 2003.

[3] J. P. Bielawski, K. A. Dunn, and Z. Yang. Rates of nucleotide substitution and mammalian nuclear gene evolution. approximate and maximum-likelihood methods lead to different conclusions. *Genetics*, 156(3):1299–1308, November 2000.

[4] L. Bromham and D. Penny. The modern molecular clock. *Nat Rev Genet*, 4:216–224, 2003.

[5] D. R. Brooks and D. A. McLennan. *Phylogeny, ecology and behaviour. A research program in comparative biology*. The University of Chicago Press, Chicago. USA, 1991.

[6] W. M. Brown, E. M. Prager, A. Wang, and A. C. Wilson. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J Mol Evol*, 18:225–239, 1982.

[7] D. A. Buonagurio, S. Nakada, W. M. Fitch, and P. Palese. Epidemiology of influenza C virus in man: multiple evolutionary lineages and low rate of change. *Virology*, 153:12–21, 1986.

[8] J. H. Camin and R. R. Sokal. A method for deducing branching sequences in phylogeny. *Evolution*, 19:311–326, 1965.

[9] L. L. Cavalli-Sforza and A. W. F. Edwards. Analysis of human evolution. In *Genetics Today. Proceeding of the XI International Congress of Genetics, The Hague, The Netherlands.*, volume 3, pages 923–933. Pergamon Press, Oxford, 1965.

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

[11] J. Chamary, J. Parmley, and L. D. Hurst. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Review in Genetics*, 7:98–108, 2006.

[12] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. A model of evolutionary change in proteins. In *Atlas of protein sequence and structure*, volume 5, pages 345–358. M. O. Dayhoff, National biomedical research foundation, Washington DC., 1978.

[13] R. W. DeBry and N. A. Slade. Cladistic analysis of restriction endonuclease cleavage maps within a maximum-likelihood framework. *Syst Zool*, 34:21–34, 1985.

[14] R. V. Eck and M. O. Dayhoff. *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Silver Spring, Maryland, 1966.

[15] R. Nielsen (ed.). *Statistical Methods in Molecular Evolution. (Statistics for Biology and Health)*. Springer-Verlag New York Inc, N.Y., first edition, 2004.

[16] B. C. Emerson, E. Paradis, and C. Thebaud. Revealing the demographic histories of species using DNA sequences. *TREE*, 16:707–716, 2001.

[17] J. S. Farris. A successive approximations approach to character weighting. *Systematics Zoology*, 18:374–385, 1969.

[18] J. S. Farris. Methods for computing Wagner trees. *Systematics Zoology*, 19:83–92, 1970.

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

[20] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17:368–376, 1981.

[21] J. Felsenstein. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genet Res*, 59:139–147, 1992.

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

[23] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. Lond. A*, 22:133–142, 1922.

[24] W. M. Fitch. Evolution of clupeine Z, a probable crossover product. *Nat New Biol*, 229:245–247, 1971.

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

[26] W. M. Fitch. Phylogenies constrained by the crossover process as illustrated by human hemoglobins and a thirteen-cycle, eleven-amino-acid repeat in human apolipoprotein A-I. *Genetics*, 86:623–644, 1977.

[27] W. M. Fitch and F. J. Ayala. The superoxide dismutase molecular clock revisited. *Proc Natl Acad Sci U S A*, 91:6802–6807, 1994.

[28] 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.

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

[30] B. Golding and J. Felsenstein. A maximum likelihood approach to the detection of selection from a phylogeny. *J Mol Evol*, 31:511–523, 1990.

[31] N. Goldman, J. P. Anderson, and A. G. Rodrigo. Likelihood-based tests of topologies in phylogenetics. *Syst Biol*, 49:652–670, 2000.

[32] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Mol Biol Evol*, 11(5):725–736, September 1994.

[33] T. Gubitz, R. S. Thorpe, and A. Malhotra. Phylogeography and natural selection in the Tenerife gecko Tarentola delalandii: testing historical and adaptive hypotheses. *Mol Ecol*, 9:1213–1221, 2000.

[34] M. S. Hafner and R. D. Page. Molecular phylogenies and host-parasite cospeciation: gophers and lice as a model system. *Philos Trans R Soc Lond B Biol Sci*, 349:77–83, 1995.

[35] D. L. Hartl and A. Clark. *Principles of polulation genetics*. Sinauer Associates, Inc., Sunderland, Massachusetts, third edition, 1997.

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

[37] P. H. Harvey and M. D. Pagel. *The comparative Method in Evolutionary Biology*. Oxford Seies in Ecology and Evolution, Oxford. England, 1991.

[38] S. B. Hedges. The origin and evolution of model organisms. *Nat Rev Genet*, 3:838–849, 2002.

[39] S. B. Hedges, H. Chen, S. Kumar, D. Y. Wang, A. S. Thompson, and H. Watanabe. A genomic timescale for the origin of eukaryotes. *BMC Evol Biol*, 1:4, 2001.

[40] M. D. Hendy and D. Penny. Branch and bound algorithm to determinate minimal evolutionary trees. *Math. Biosci.*, 60:309–368, 1982.

[41] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89:10915–10919, 1992.

[42] W. Hennig. *Grundzüge einer theorie der phylogenetischen systematik*. Deutscher Zentralverlag, Berlin, 1950.

[43] W. Hennig. *Phylogenetic systematics*. University of Illinois Press, Urbana, 1966.

[44] J. Hey. The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. *Genetics*, 128:831–840, 1991.

[45] D. M. Hillis and J. P. Huelsenbeck. Support for dental HIV transmission. *Nature*, 369:24–25, 1994.

[46] M. Holder and P. O. Lewis. Phylogeny estimation: traditional and Bayesian approaches. *Nat Rev Genet*, 4:275–284, 2003.

[47] J. P. Huelsenbeck, F. Ronquist, R. Nielsen, and J. P. Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294:2310–2314, 2001.

[48] Y. Ina. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J Mol Evol*, 40(2):190–226, February 1995.

[49] D. T. Jones, W. R. Taylor, and J. M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, 8:275–282, 1992.

[50] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In M. N. Munro, editor, *Mammalian protein metabolism*, volume III, pages 21–132. Academic Press, N. Y., 1969.

[51] K. K. Kidd and L. A. Sgaramella-Zonta. Phylogenetic analysis: concepts and methods. *Am J Hum Genet*, 23:235–252, 1971.

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

[53] H. Kishino and M. Hasegawa. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J Mol Evol*, 29:170–179, 1989.

[54] A. G. Kluge and J. S. Farris. Quantitative phyletics and the evolution of anurans. *Systematics Zoology*, 18:1–36, 1969.

[55] S. Kumar and S. B. Hedges. A molecular timescale for vertebrate evolution. *Nature*, 392:917–920, 1998.

[56] A. Kurosky, D. R. Barnett, T. H. Lee, B. Touchstone, R. E. Hay, M. S. Arnott, B. H. Bowman, and W. M. Fitch. Covalent structure of human haptoglobin: a serine protease homolog. *Proc Natl Acad Sci U S A*, 77:3388–3392, 1980.

[57] P. O. Lewis. Phylogenetic systematics turns over a new leaf. *TRENDS IN ECOLOGY AND EVOLUTION*, 16:30–37, 2001.

[58] W. H. Li. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J Mol Evol*, 36(1):96–99, January 1993.

[59] W. H. Li, C. I. Wu, and C. C. Luo. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol*, 2(2):150–174, March 1985.

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

[61] P. Lio and N. Goldman. Models of molecular evolution and phylogeny. *Genome Res*, 8:1233–1244, 1998.

[62] P. Lio and N. Goldman. Using protein structural information in evolutionary inference: transmembrane proteins. *Mol Biol Evol*, 16:1696–1710, 1999.

[63] P. Lio and N. Goldman. Modeling mitochondrial protein evolution using structural information. *J Mol Evol*, 54:519–529, 2002.

[64] E. P. Martins. *Phylogenies and the comparative method in animal behavior*. Oxford University Press, Oxford. England, 1996.

[65] E. Mayr. *Principles of systematics zoology*. McGraw-Hill, New York, 1969.

[66] E. Mayr. *The growth of biological thought. Diversity, evolution and inheritance*. Belknap-Harvard, Massachusetts, 1982.

[67] A. Meyer. Hox gene variation and evolution. *Nature*, 391:225, 227–8, 1998.

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

[69] T. Miyata and T. Yasunaga. Molecular evolution of mrna: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J Mol Evol*, 16(1):23–36, September 1980.

[70] C. Moritz. Strategies to protect biological diversity and the evolutionary processes that sustain it. *Syst Biol*, 51:238–254, 2002.

[71] T. Muller and M. Vingron. Modeling amino acid replacement. *J Comput Biol*, 7:761–776, 2000.

[72] Galtier N., O. Gascuel, and A. Jean-Marie. Markov models in molecular evolution. In R. Nielsen, editor, *Statistical Methods in Molecular Evolution. (Statistics for Biology and Health)*. Springer-Verlag New York Inc, N.Y., first edition, 2004.

[73] M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*, 3(5):418–426, September 1986.

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

[75] R. D. Page, R. H. Cruickshank, M. Dickens, R. W. Furness, M. Kennedy, R. L. Palma, and V. S. Smith. Phylogeny of Philoceanus complex seabird lice (Phthiraptera: Ischnocera) inferred from mitochondrial DNA sequences. *Mol Phylogenet Evol*, 30:633–652, 2004.

[76] R. D. M. Page. *Tangled trees.* The University of Chicago Press, Chicago, London, 2001.

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

[78] P. Pamilo and N. O. Bianchi. Evolution of the zfx and zfy genes: rates and interdependence between the genes. *Mol Biol Evol*, 10(2):271–281, March 1993.

[79] A. L. Panchen. Richard Owen and the homology concept. In Brian K. Hall, editor, *Homology. The hierarchical basis of comparative biology*, pages 21–62. Academic Press, N. Y., 1994.

[80] D. Posada. Selecting models of evolution. Theory and practice. In M. Salemi and A. M. Vandamme, editors, *The phylogenetic handbook. A practical approach to DNA and protein phylogeny*, pages 256–282. Cambridge University Press, UK, 2003.

[81] D. Posada and K. A. Crandall. MODELTEST: testing the model of DNA substitution. *Bioinformatics*, 14:817–818, 1998.

[82] D. Posada and K. A. Crandall. Selecting the best-fit model of nucleotide substitution. *Syst Biol*, 50:580–601, 2001.

[83] J. Raymond, J. L. Siefert, C. R. Staples, and R. E. Blankenship. The natural history of nitrogen fixation. *Mol Biol Evol*, 21:541–554, 2004.

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

[85] S. Rudikoff, W. M. Fitch, and M. Heller. Exon-specific gene correction (conversion) during short evolutionary periods: homogenization in a two-gene family encoding the beta-chain constant region of the T-lymphocyte antigen receptor. *Mol Biol Evol*, 9:14–26, 1992.

[86] A. Rzhetsky and M. Nei. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *J Mol Evol*, 35:367–375, 1992.

[87] A. Rzhetsky and M. Nei. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol Biol Evol*, 10:1073–1095, 1993.

[88] A. Rzhetsky and M. Nei. METREE: a program package for inferring and testing minimum-evolution trees. *Comput Appl Biosci*, 10:409–412, 1994.

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

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

[91] H. A. Schmidt, K. Strimmer, M. Vingron, and A. von Haeseler. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 18:502–504, 2002.

[92] C. Scholtissek, S. Ludwig, and W. M. Fitch. Analysis of influenza A virus nucleoproteins for the assessment of molecular genetic mechanisms leading to new phylogenetic virus lineages. *Arch Virol*, 131:237–250, 1993.

[93] H. Shimodaira and M. Hasegawa. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol*, 16:1114–1116, 1999.

[94] G. G. Simpsom. *Principles of animal taxonomy.* Columbia University Press, New York, 1961.

[95] M. Slatkin and W. P. Maddison. A cladistic measure of gene flow inferred from the phylogenies of alleles. *Genetics*, 123:603–613, 1989.

[96] P. Sneath. The application of computers to taxonomy. *Journal of general microbiology*, 17:201–226, 1957.

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

[98] K. Strimmer and A. Rambaut. Inferring confidence sets of possibly misspecified gene trees. *Proc R Soc Lond B Biol Sci*, 269:137–142, 2002.

[99] Y. Surget-Groba, B. Heulin, C. P. Guillaume, R. S. Thorpe, L. Kupriyanova, N. Vogrin, R. Maslak, S. Mazzotti, M. Venczel, I. Ghira, G. Odierna, O. Leontyeva, J. C. Monney, and N. Smith. Intraspecific phylogeography of Lacerta vivipara and the evolution of viviparity. *Mol Phylogenet Evol*, 18:449–459, 2001.

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

[101] 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.

[102] 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.

[103] W. H. Jr. Wagner. Problems in the classifications of ferns. In *Recent Advances in Botany. IX International Botanical Congress. Montreal*, pages 841–844, Toronto, 1959. University of Toronto Press.

[104] S. Whelan and N. Goldman. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol*, 18:691–699, 2001.

[105] S. Whelan, P. Lio, and N. Goldman. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet*, 17:262–272, 2001.

[106] 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.

[107] Z. Yang. Among-site variation and its inpact on phylogenetic analises. *TREE*, 11:367–371, 1996.

[108] Z. Yang. Paml: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13(5):555–556, October 1997.

[109] Z. Yang. Adaptive Molecular Evolution. In D. Balding, M. Bishop, and C. Cannings (eds.), editors, *Handbook of Statistical Genetics*. Wiley J. and Sons Ltd., N.Y., second edition, 2003.

[110] Z. Yang and R. Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*, 17(1):32–43, January 2000.

[111] S. H. Yeh, H. Y. Wang, C. Y. Tsai, C. L. Kao, J. Y. Yang, H. W. Liu, I. J. Su, S. F. Tsai, D. S. Chen, and P. J. Chen. Characterization of severe

acute respiratory syndrome coronavirus genomes in Taiwan: molecular epidemiology and genome evolution. *Proc Natl Acad Sci U S A*, 101:2542–2547, 2004.

[112] R. Nielsen Z. Yang. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J Mol Evol.*, 46:409–418, 1998.

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