

Introduction to Phylogenetic Inference

“Our classifications will come to be, as far as they can be so made, genealogies” Darwin 1887

The origin of phylogenetic systematics can be traced in **Willi Hennig’s** 1950 book (Hennig 1966). Phylogenetic systematics is based on **recency of common descent (genealogy)** and its goal is to produce hypotheses of genealogical relationships among monophyletic groups of organisms.

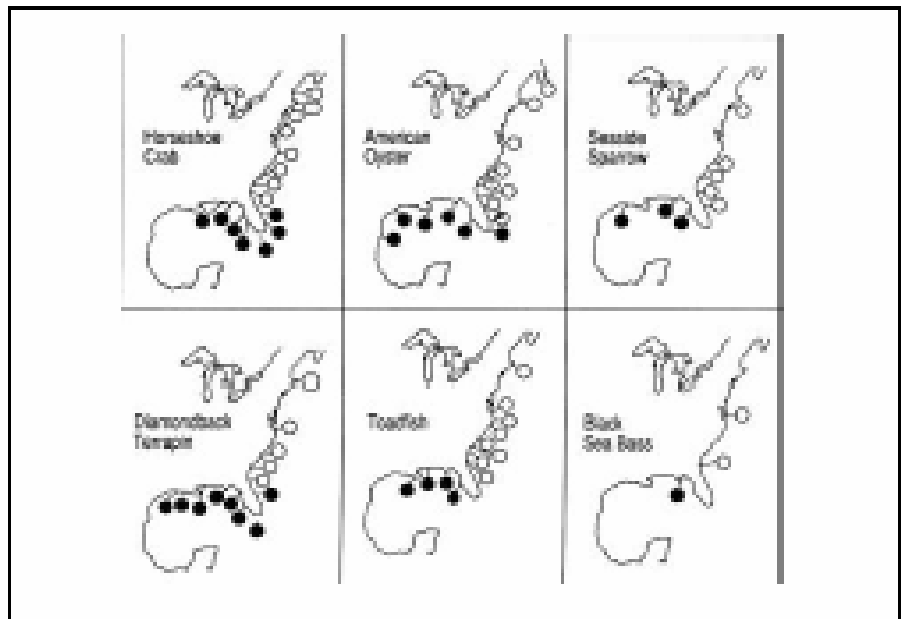
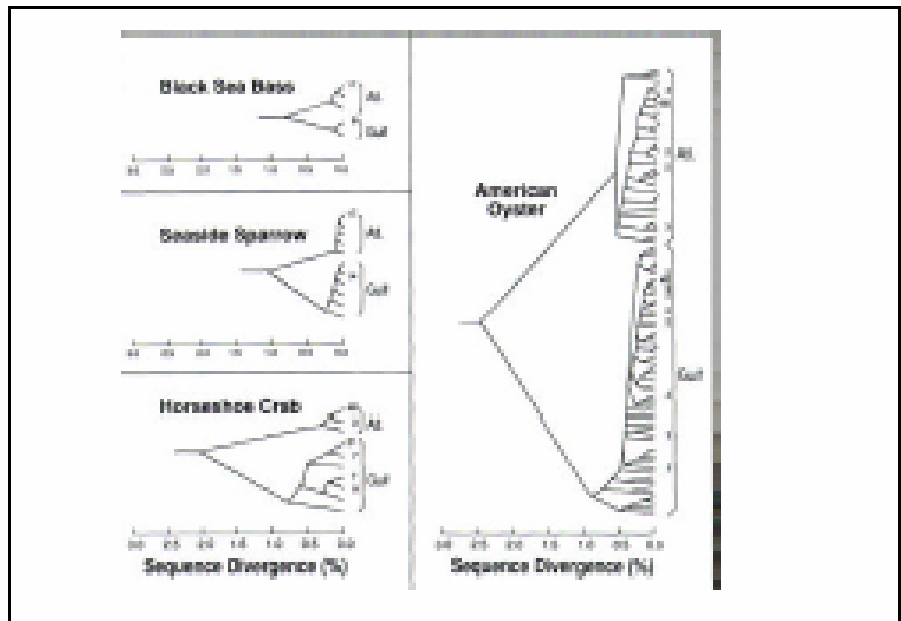
Why do we need phylogenies? What can they tell us?

1) Intrinsic of phylogeny: to reconstruct the true relationships between organisms

2) They are powerful tools for the analysis of any type of information affected by history (i.e. all of biology, for a start)

-Phylogenies are no longer exclusive to ecology, systematics & biogeography but also to molecular biologists, physiologists, neuroscientists, developmental biologists and informaticists.

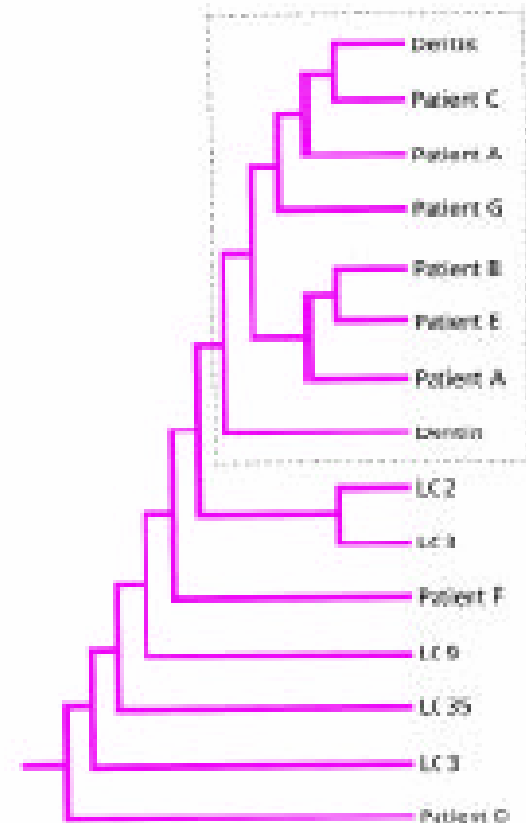
Biogeography of species distributed in the Atlantic and Gulf of Mexico



Phylogenetic trees used in forensic science.

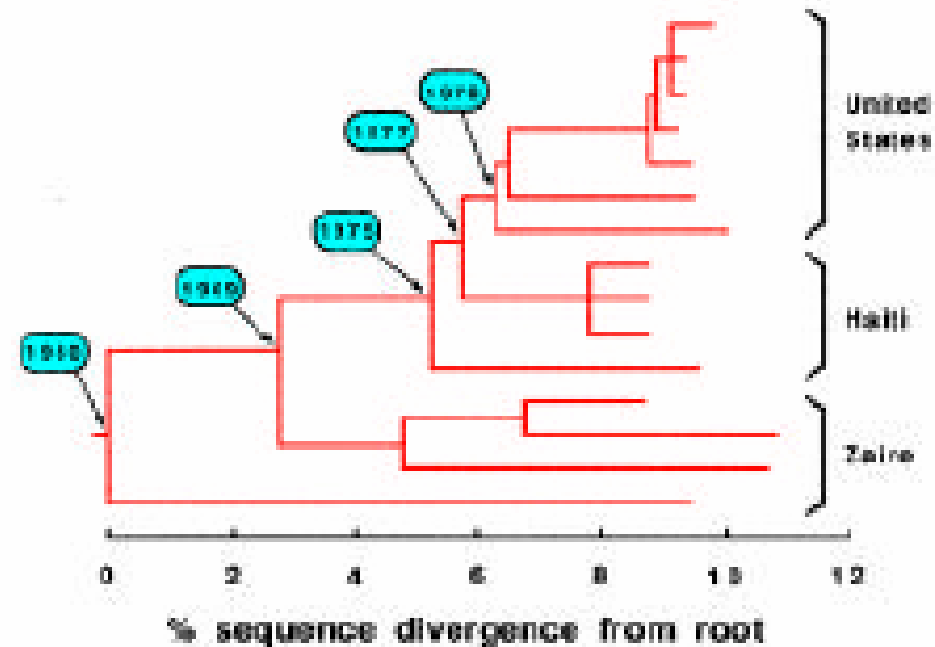
Florida dentist case: (Ou et al.1992)

- Dentist test +ve for HIV
- 7 of his patients also tested +ve.
- DNA sequence comparisons conducted for dentist, 7 infected patients and 35 other HIV carriers from local area.
- Results: HIV from 5 of 7 patients closely related to that of dentist
- First genetic confirmation of HIV transmission from an infected healthworker to clients.

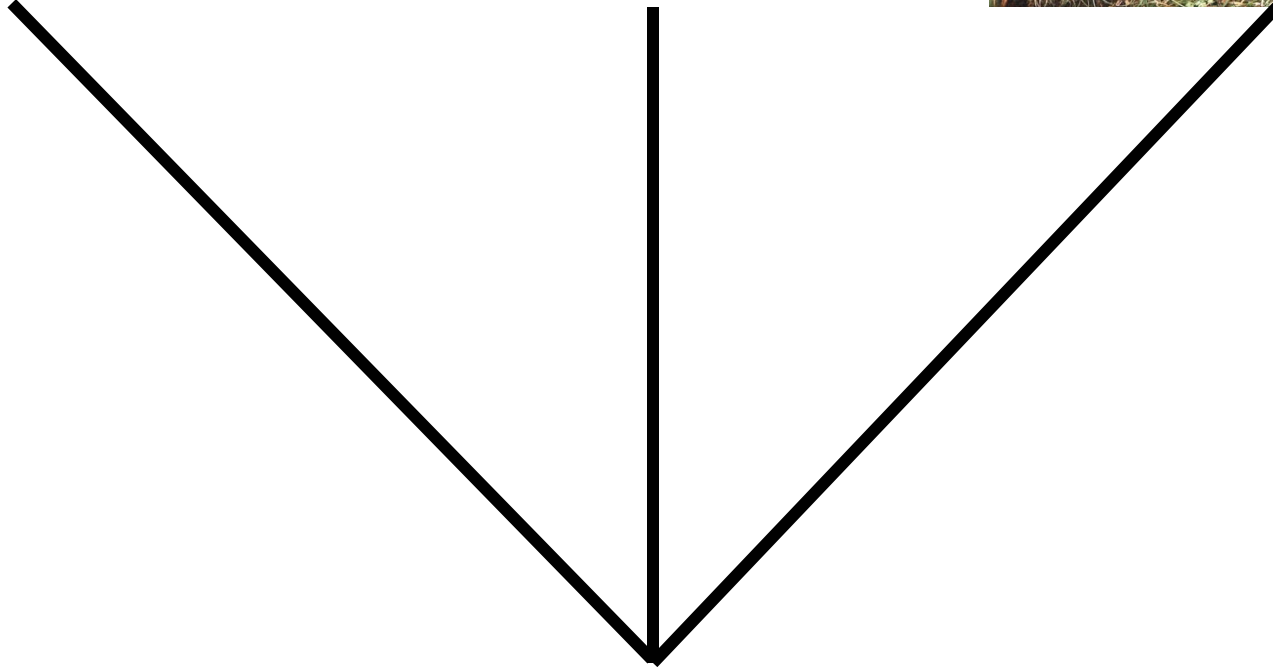


Biogeographic epidemiology of HIV

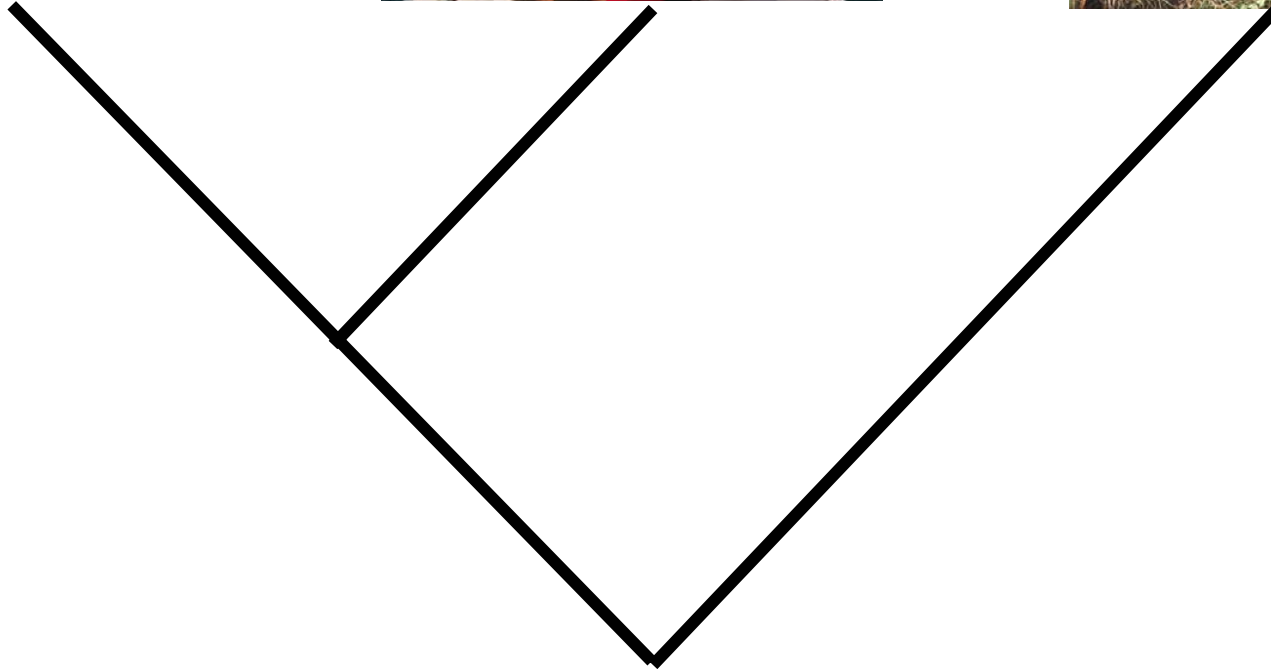
- W.H. Li et al. (1988) analysed sequences from 15 HIV isolates from US, Haiti and Africa
- Results consistent w/ african ancestry of HIV, its subsequent spread to Haiti and on to U.S.A.
- Li et al used estimates of nucleotide substitution rate to estimate divergences.
- Inferred that separation from Africa to New World was 1969; from Haiti to US was 1977!



Phylogeny of higher Apes

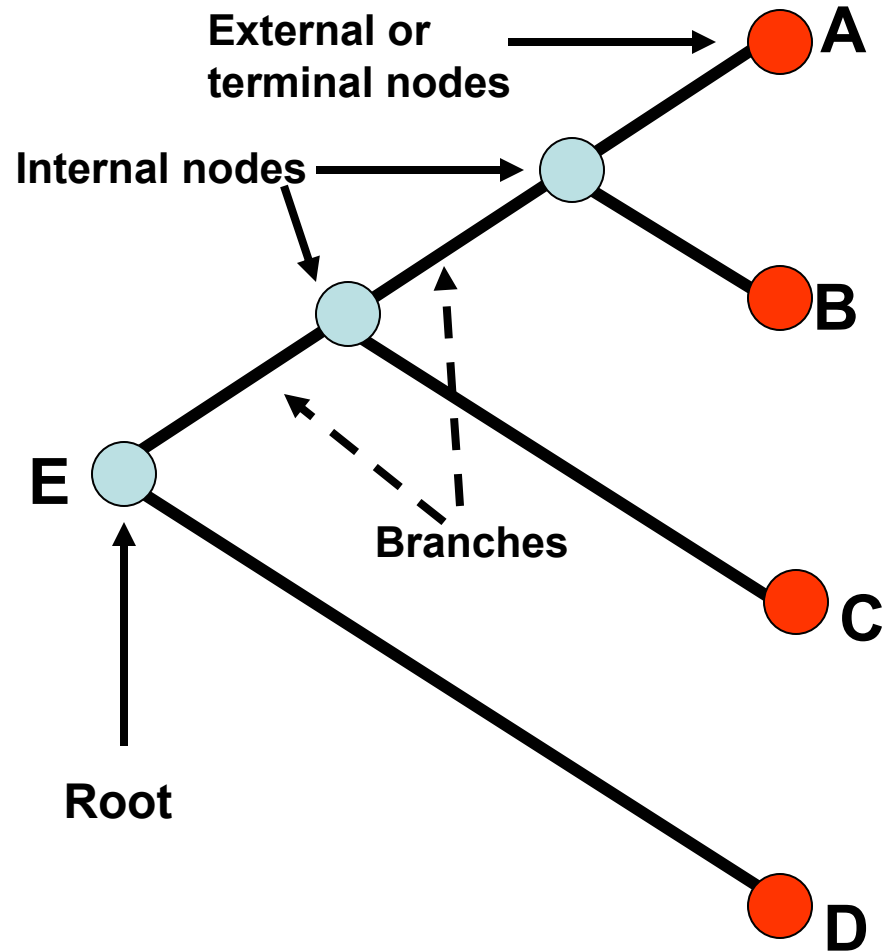


Phylogeny of higher Apes



What is a phylogenetic tree?

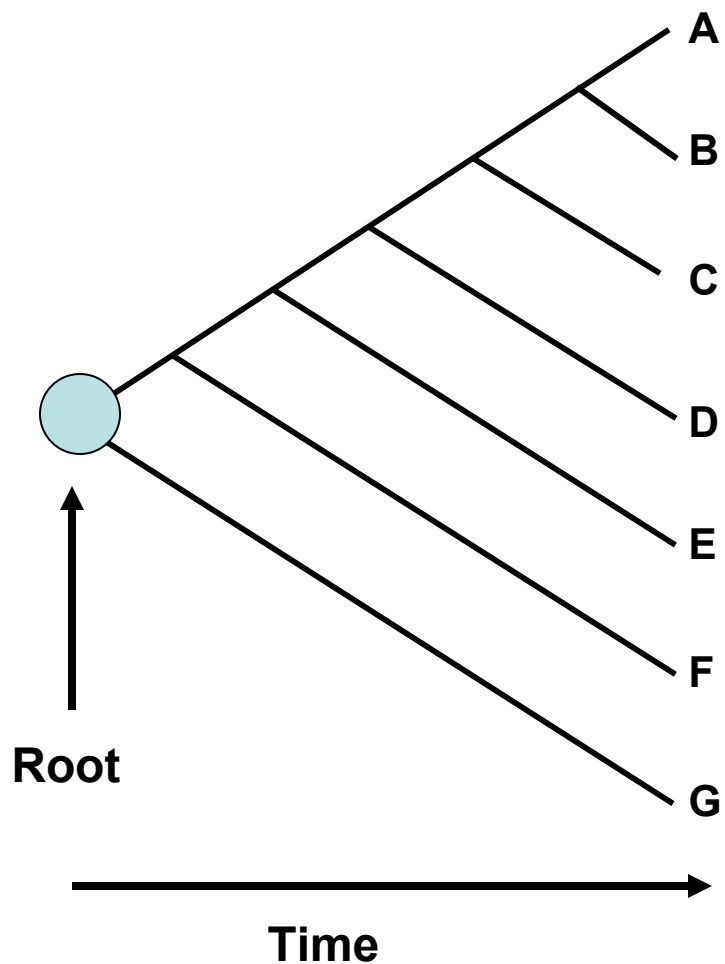
“A graph depicting the ancestor-descendant relationships between organisms or gene sequences. The sequences are the tips of the tree. Branches of the tree connect the tips to their (unobservable=hypothetical) ancestral sequences” in (Holder and Lewis 2003)



A-D = Operational Taxonomic Units (OTUs)
E = hypothetical ancestor for all OTUs

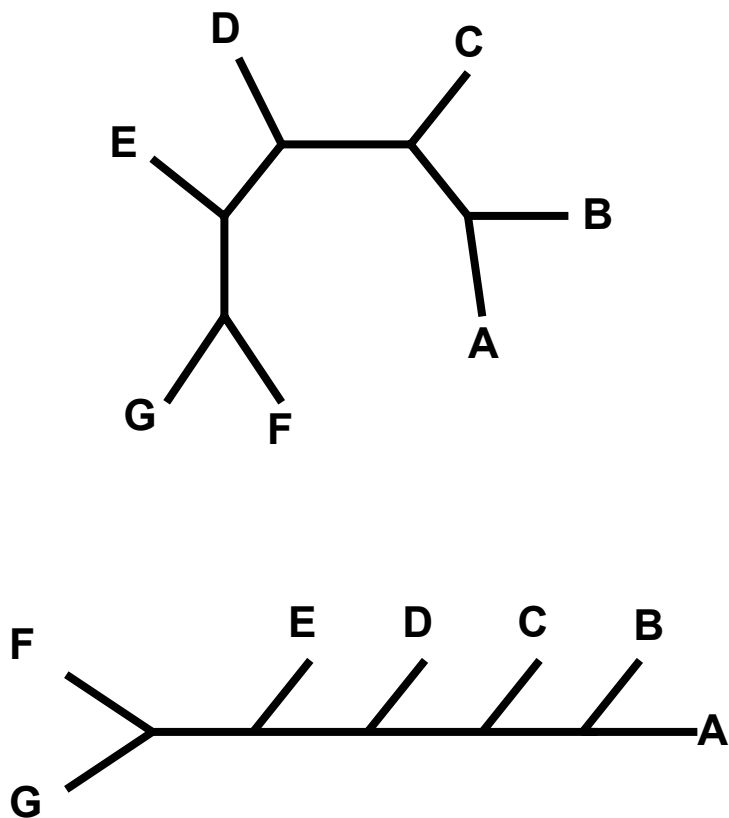
Rooted trees

They have direction corresponding to evolutionary time

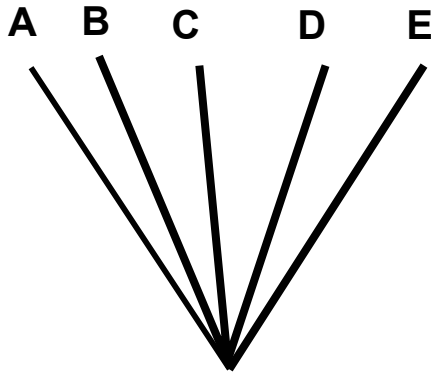


vs. Unrooted trees

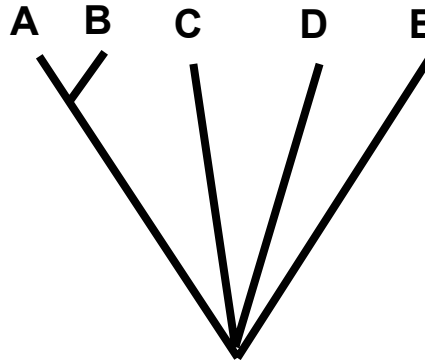
We cannot talk about of ancestors and descendants



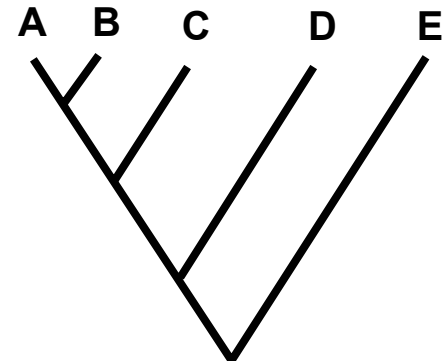
Star Tree



Partially Resolved Tree

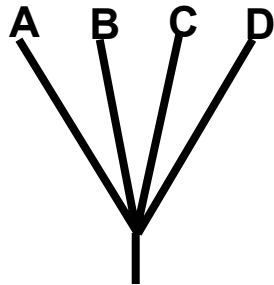


Resolved Tree

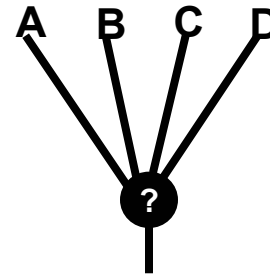


Hard vs. Soft Polytomy

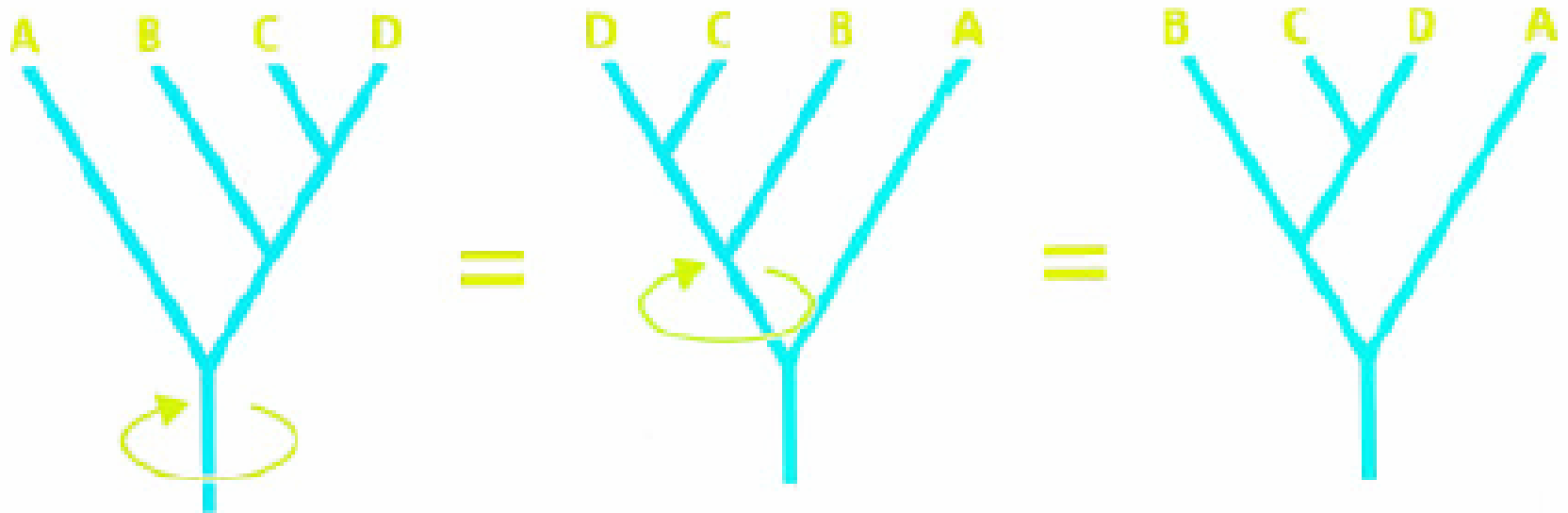
Hard Polytomy
Ancestral lineage splits “simultaneously”
into multiple lineages



Soft Polytomy
Little or no statistical support
for complete tree resolution

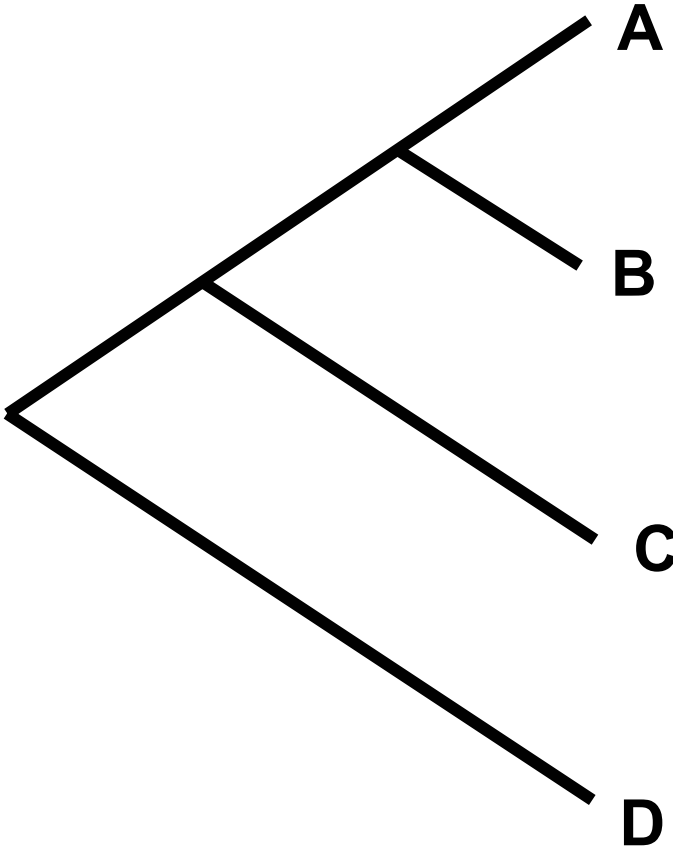


Trees are like mobiles

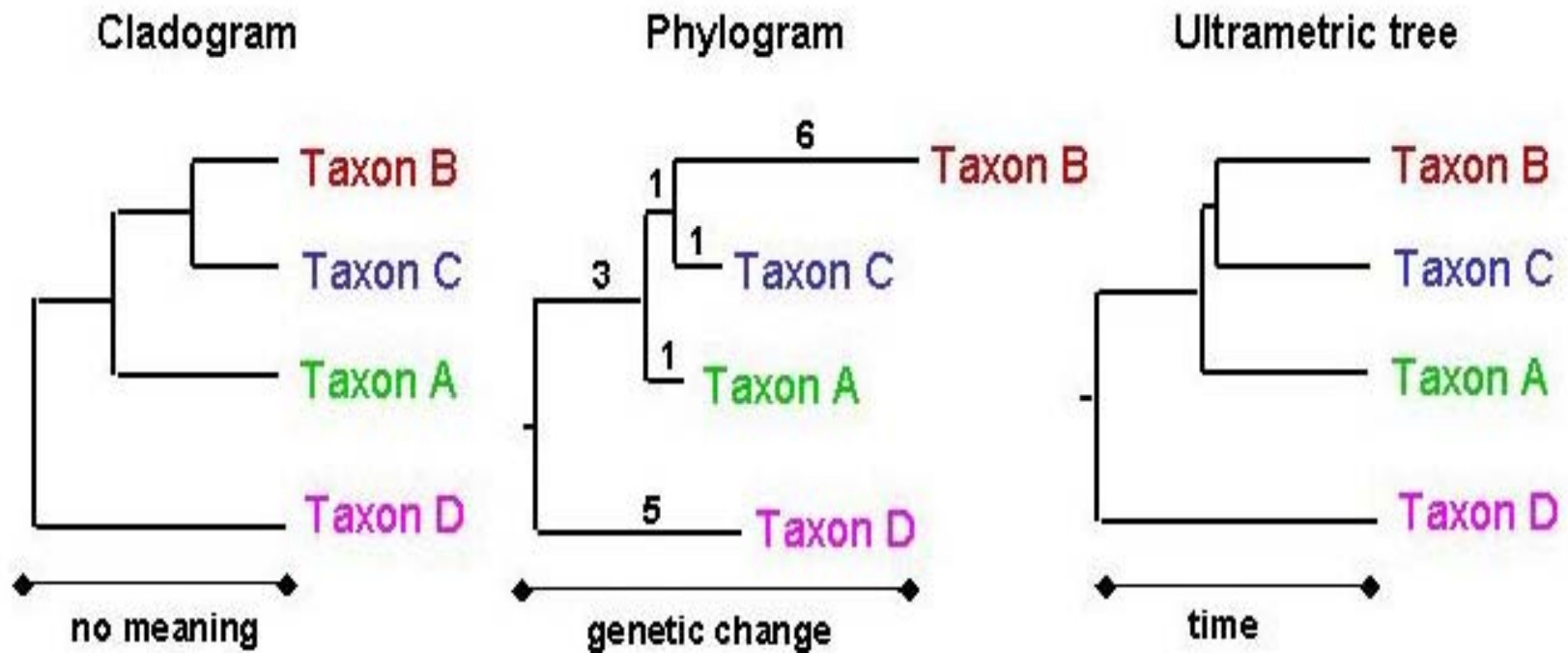


all 3 trees depict same topological information

Parenthetic Notation



$((A,B), C), D$ –Same information as in the tree



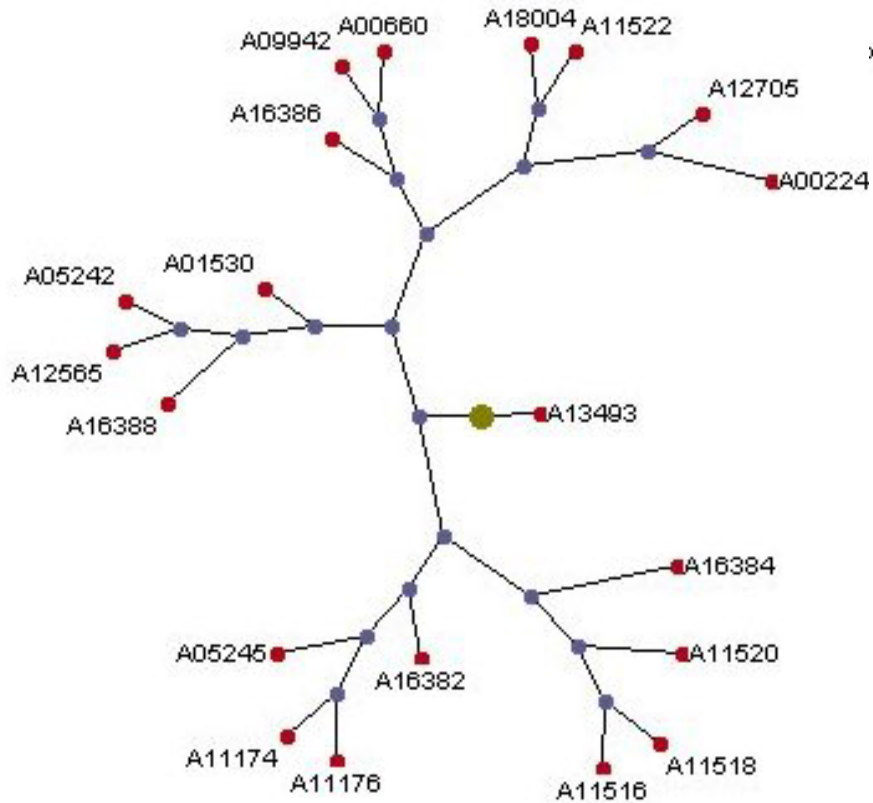
A **cladogram** is a branching diagram representing the most parsimonious distribution of shared derived characters (synapomorphies) within a set of taxa. The branching pattern of a cladogram shows the relative relationships among taxa, it is not a true "evolutionary tree" of how those relationships came to be.

Phylogram is a phylogenetic tree that indicates the relationships between the taxa and also conveys a sense of time or rate of evolution.

An **ultrametric tree** is a rooted tree where all tips are equidistant from the root. The branch lengths of an ultrametric tree are proportional to the divergent time (i.e. implying a constant rate of evolution).

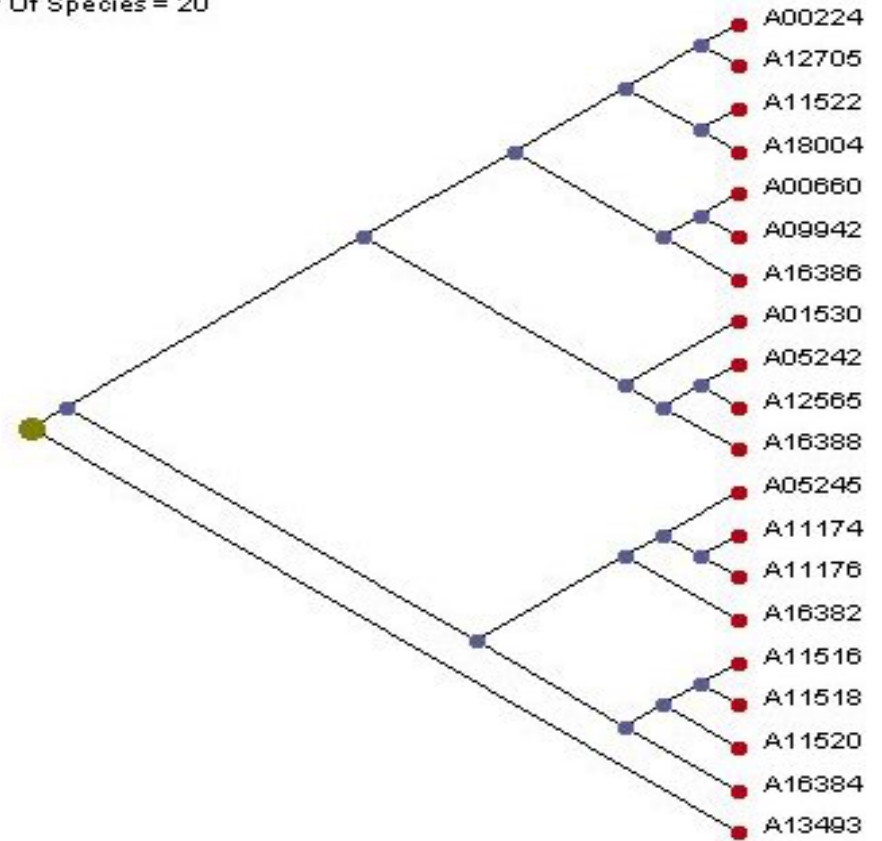
More examples of phylogenetic trees

A) Radial tree

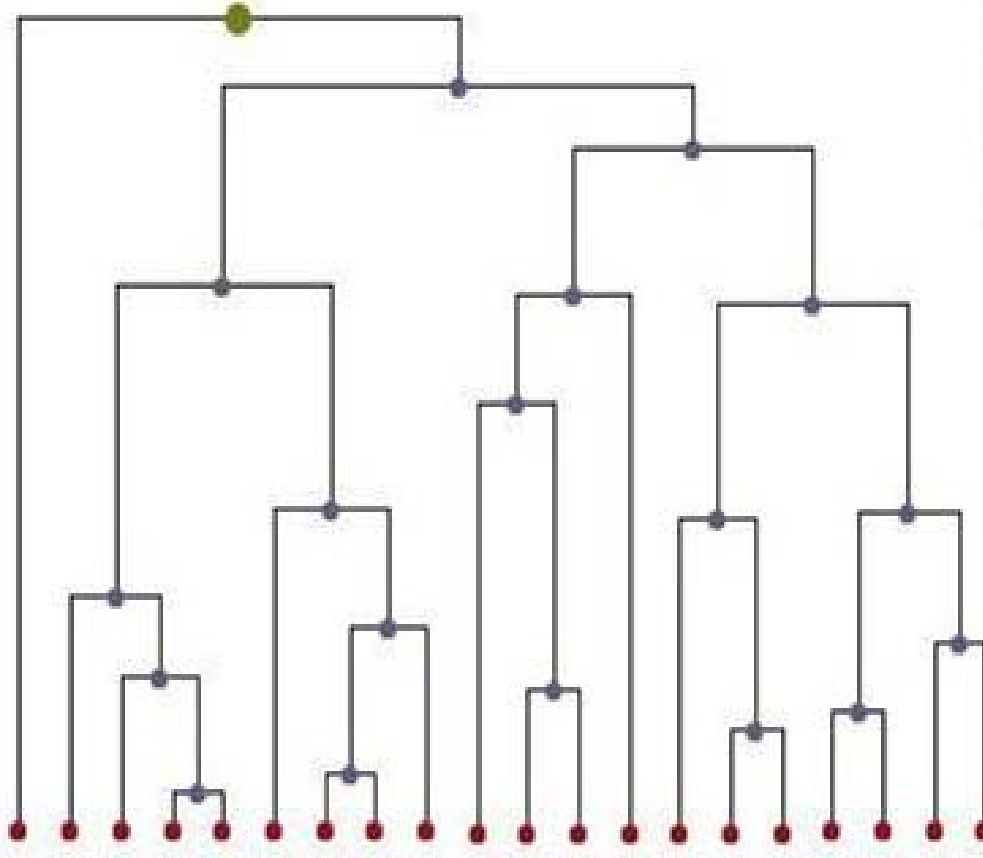


B) Slanted cladogram

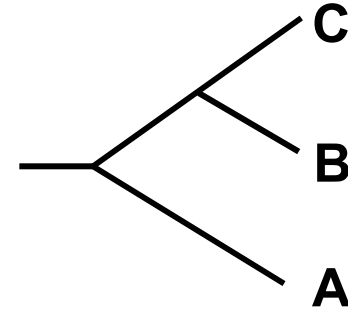
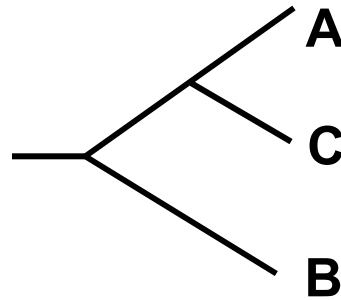
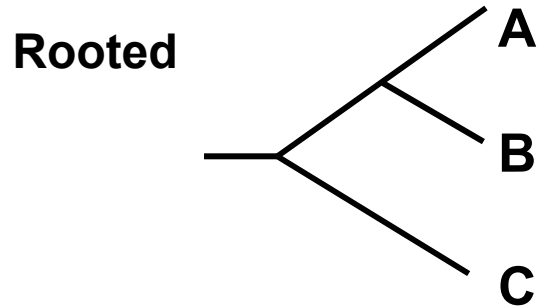
Number Of Species = 20



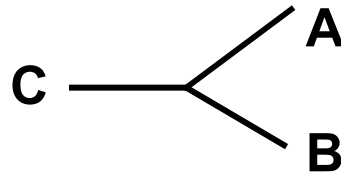
C) Coalescence tree



Possible topologies of rooted (3) vs. unrooted (1) trees for 3 taxa (OTUs)



Unrooted



Number of OTUs	Number of rooted trees	Number of unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
..
9	2,027,025	135,135
10	34,459,425	2,027,025

Read (Felsenstein 1978) for more information on the number of phylogenetic trees. For 20 sequences 8,200,794,532,637,891,559,000 trees. For a recent study of 135 human mtDNA sequences 2.113×10^{267} trees.

Phylogenetic Methods

Molecular phylogenetics is the study of evolutionary relationships among organisms or genes by a combination of molecular biology and statistical techniques” (Li 1997)

A) Distance Matrix methods: UPGMA (Sneath and Sokal 1973), Minimum Evolution (Rzhetsky and Nei 1992), Neighbor-Joining (Saitou and Nei 1987),

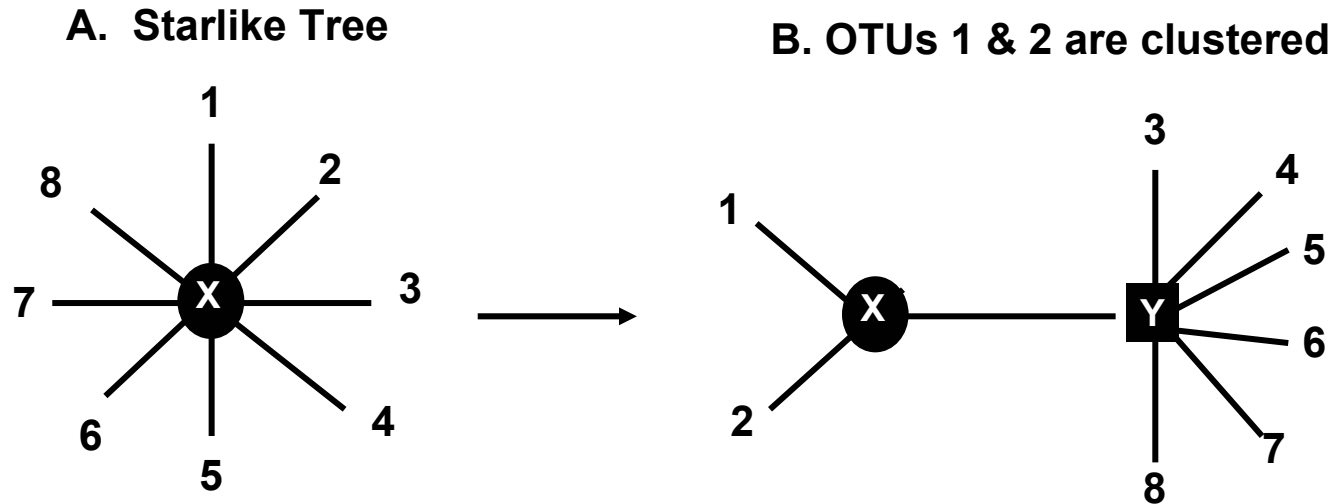
Unweighted Pair-Group Method with Arithmetic Mean (UPGMA). This is the simplest method for tree reconstruction but at the same time phylogeneticists have long abandoned this method. If the DNA substitution rates are approximately constant across lineages it works well (but this is an assumption, violated in the majority of data sets). In this case, genetic distance among taxa has a linear relationship with divergence.

Minimum Evolution (ME). For all possible topologies, estimate the length of each branch from the estimated pairwise distance between taxa and compute the sum (S) of all branch-length estimates. The **minimum evolution optimality criterion** is to choose the trees with the smallest S value.

Phylogenetic Methods

Neighbor-Joining (NJ) is a fast, approximate method for finding the minimum-evolution tree (i.e. the tree with the smallest S). NJ is a distance matrix method producing an unrooted tree without the assumption of a clock. NJ keeps track of nodes on the tree as opposed to taxa or clusters of taxa. NJ algorithm involves construction of a modified distance matrix in which the separation between each pair of nodes is adjusted on the basis of their average divergence from all other nodes. NJ is computationally fast. Opponents suggest that the NJ tree is a good starting point, but not the endpoint.

N-J algorithm



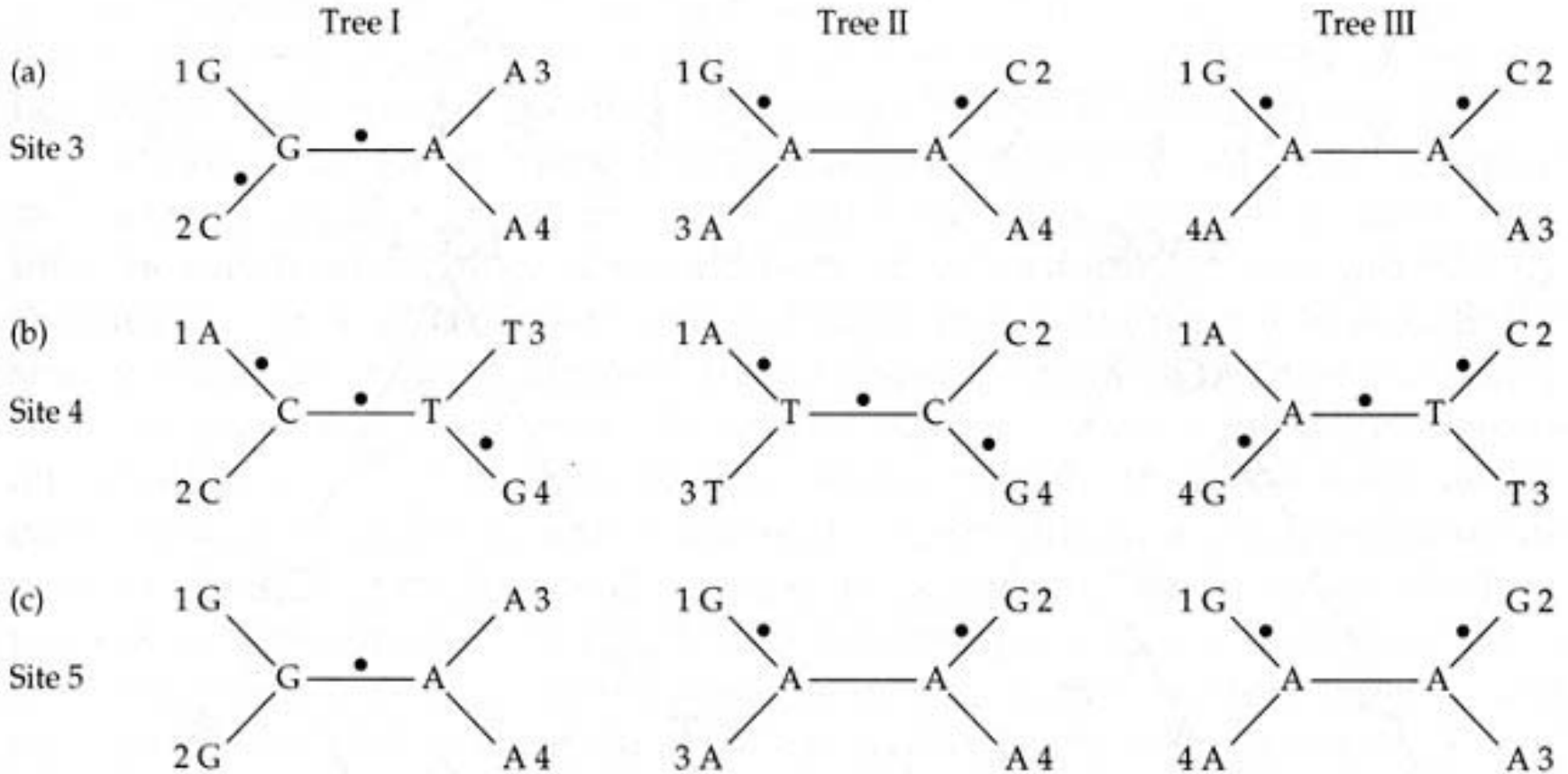
B) Maximum Parsimony Methods

The **maximum parsimony optimality criterion** (i.e. Occam's Razor: one should not increase, beyond what is necessary, the number of entities required to explain anything) searches for a tree that requires the smallest number of evolutionary changes given the data. The tree with the least changes (steps) is the **maximum parsimony tree**. Often more than one tree with the same minimum number of changes is found, so that no unique tree can be inferred. The method discussed below was first developed for amino acid sequence data (Eck and Dayhoff 1966) and was later modified for use on nucleotide sequences (Fitch 1971).

Example from (Li 1997):

Sequence	1	2	3	4	5	6	7	8
9								
1	T	A	G	A	G	A	G	C
2	T	G	C	C	G	A	G	C
3	T	G	A	T	A	A	C	C
4	T	G	A	G	A	A	C	C

A nucleotide site is phylogenetically informative only if it favors some trees over the others. Sites 5,7, and 9 are parsimony informative sites. All others are constant (Sites 1, 6, and 8) and uninformative (Sites 2, 3, and 4).



Bayesian Method

- A branch of statistics that focuses on the posterior probability of hypotheses.
- Posterior probabilities are estimated, based on some model (prior expectations), after learning something about the data. The posterior probability is proportional to the product of the prior probability and the likelihood.
- The Bayesian method is similar to the maximum, likelihood method in that a model of nucleotide substitution has to be postulated and the “best” trees are consistent with both the model and the data.
- It differs from ML since Bayesian analysis seek the tree that maximizes the probability of the tree given the data and the model of evolution (ML seeks the tree that maximizes the probability of observing the data given the tree).

Useful papers:

(Rannala and Yang 1996; Yang and Rannala 1997; Huelsenbeck and Ronquist 2001; Huelsenbeck et al. 2001).

Assessing confidence of phylogenetic trees (or how strongly does the data support the species or gene relationships depicted on the tree)-Bootstrapping (Felsenstein 1985). Alternative methods exist such as Bremer support in (Bremer 1988)

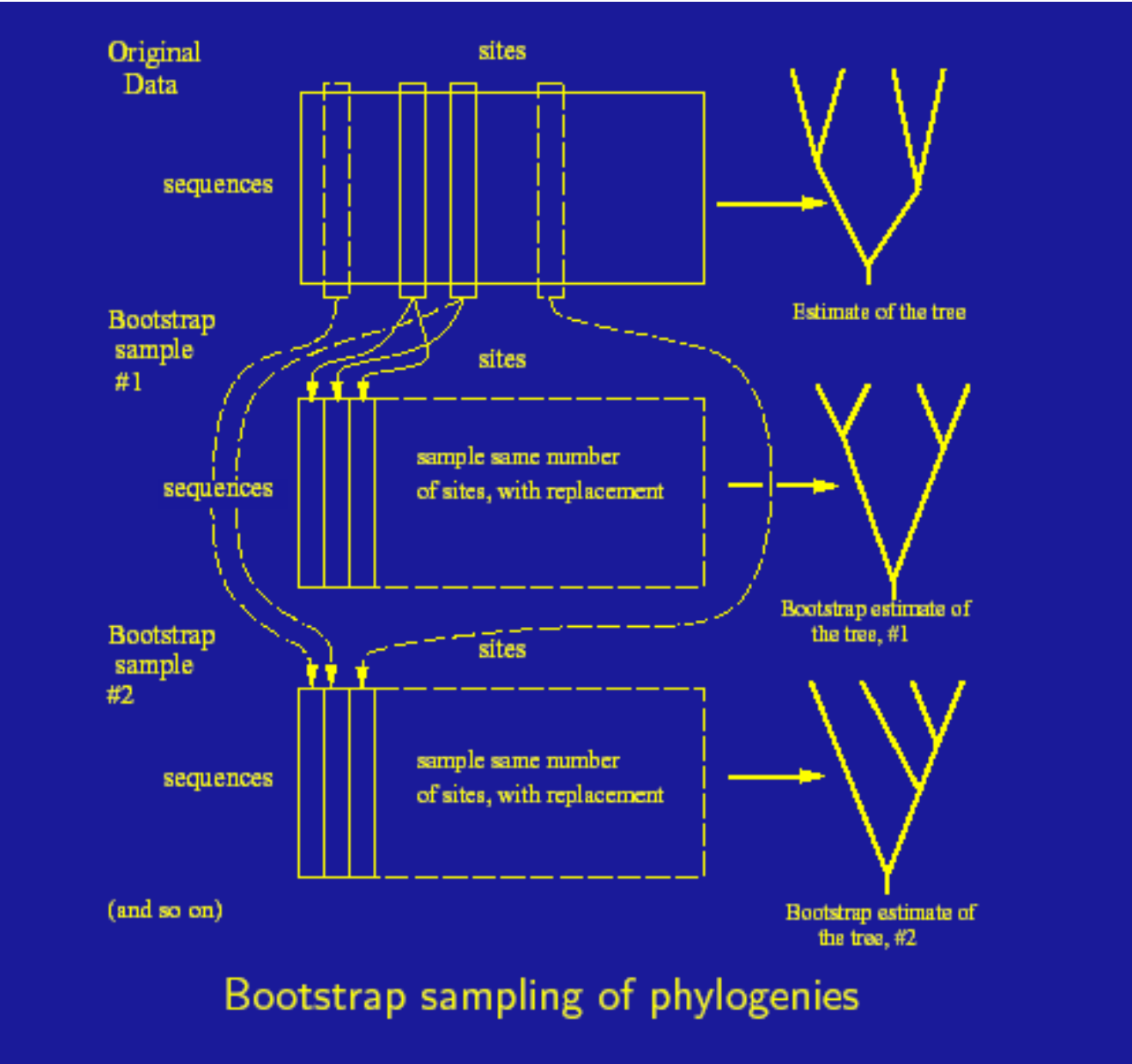


Table 1. Comparison of phylogenetic methods^{1,2} (from (Holder and Lewis 2003)

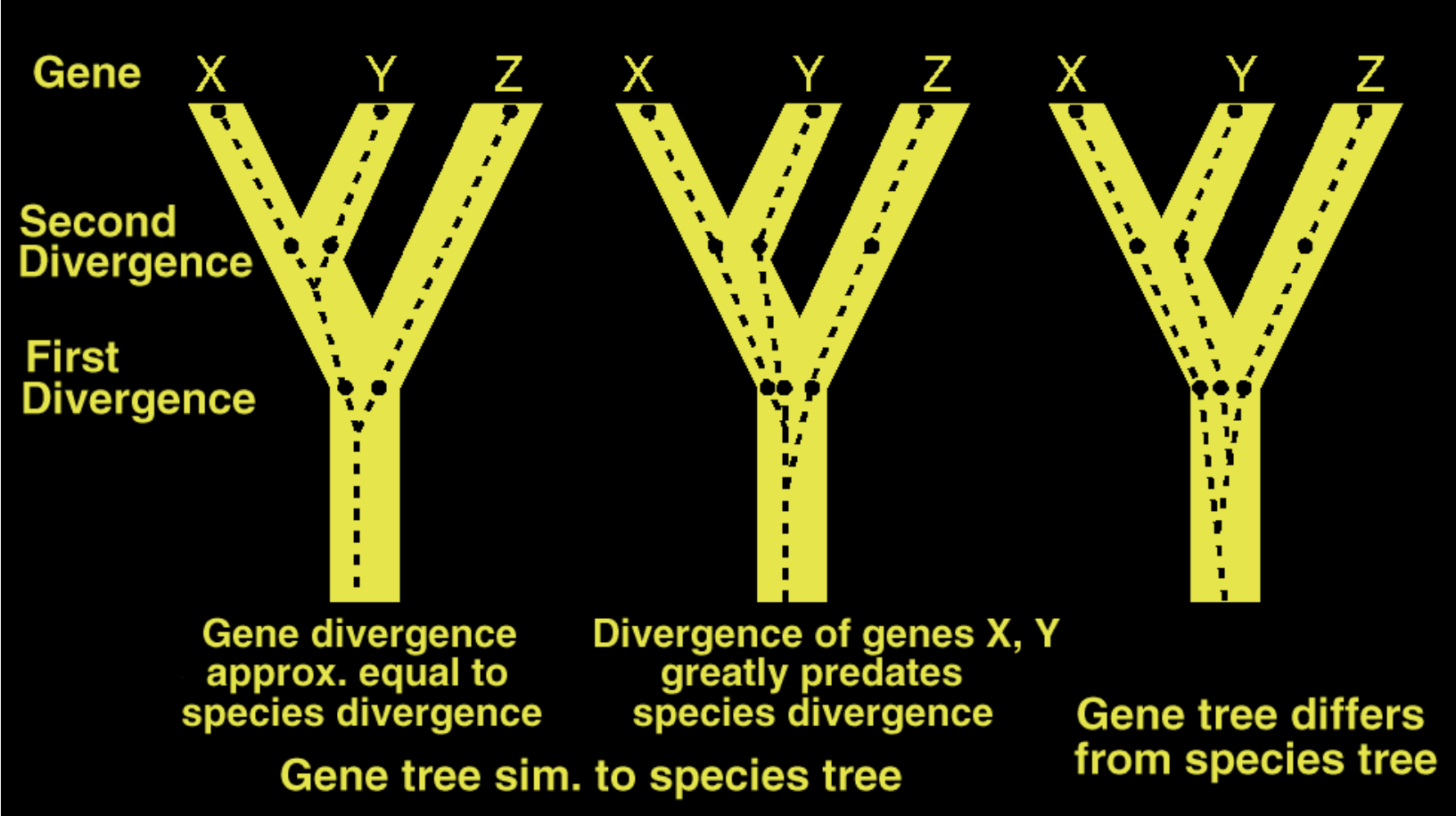
Method	Advantages	Disadvantages	Software
NJ	Fast	Sequences are transformed to Distances and some info is lost Reliable estimates of pairwise distance hard to obtain for divergent sequences	PAUP* MEGA PHYLIP
Parsimony	Fast enough for the analysis of 100s of sequences. Robust if branches are short (closely related sequences or dense taxon sampling)	Poor performance when variation in branch length.	PAUP* NONA MEGA PHYLIP
Min Evol	Uses models to correct for unseen changes	Poor performance when genetic distances are large	PAUP* MEGA PHYLIP
Max Lik	The likelihood fully captures what the data tell us about the phylogeny under a given model	Computationally very slow	PHYLIP PAUP* MEGA
Bayesian	Prior distributions must be specified It could be provide a faster way to assess support for trees than max. lik. bootstrapping	Difficult to determine if the Markov chain Monte Carlo (MCMC) approximation has run long enough	BAMBE MrBayes

1For a quick guide to aid beginners in getting started creating phylogenetic trees look (Hall 2001)

**2For the most comprehensive list of phylogeny programs look at
<http://evolution.genetics.washington.edu/phylip/software.html>**

Species trees vs. Gene trees

All the phylogenetic trees we will encounter are based on 1 or more genes. It is assumed that the chosen genes capture the “true” organismal history of the taxa. There are several examples that show that gene trees are not species trees.



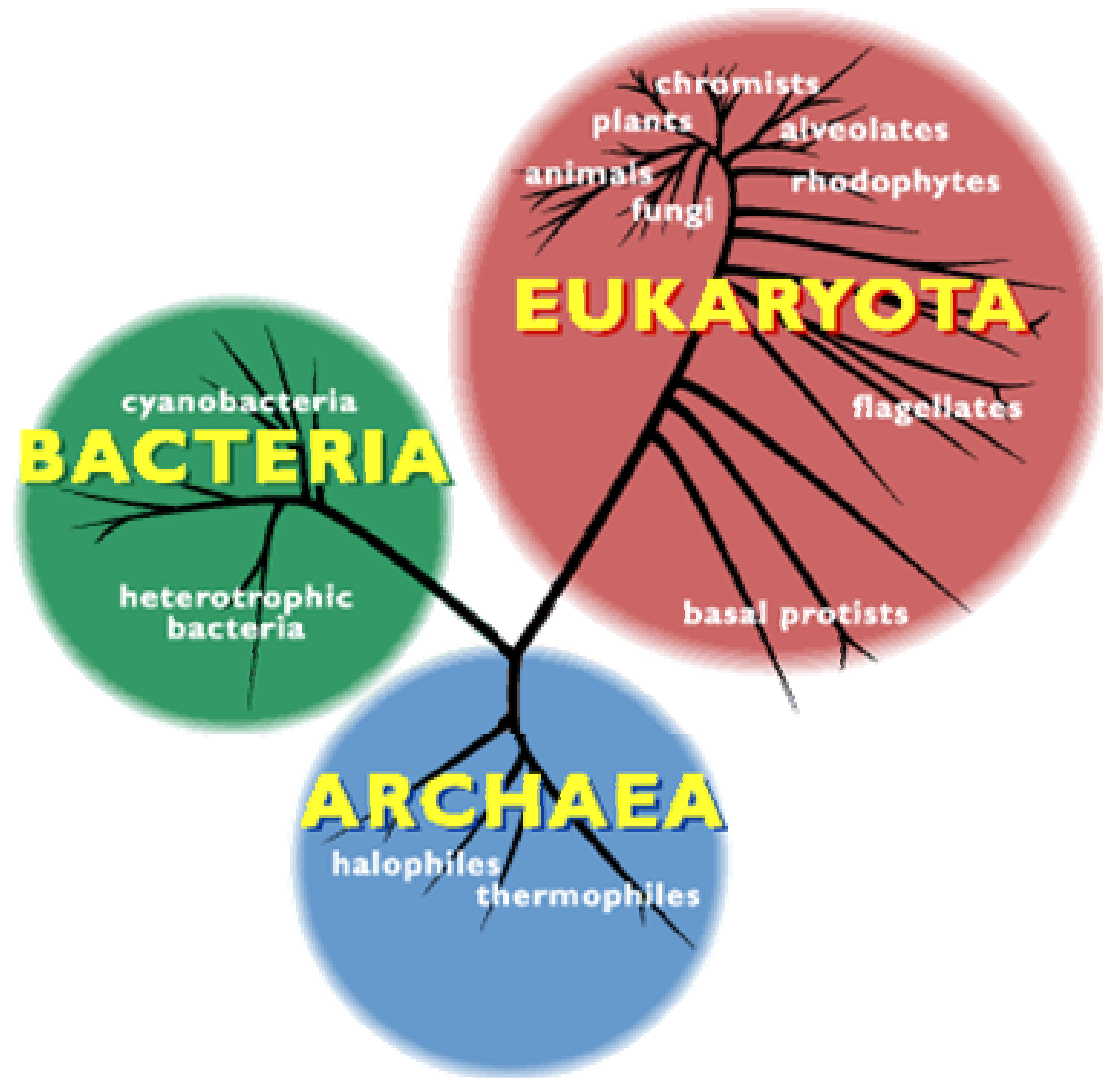
Additional reading: (Hudson 1983; Pamilo and Nei 1988; Avise 1989; Nichols 2001)

Microbiologists use ribosomal DNA (16S) to identify and name species of bacteria.

Microbial ancestries would be difficult, if not impossible, to untangle because of horizontal transfer

In the journal *Science* (Aug 8 2003 issue) Daubin et al. examined sequences of related bacteria and identified a number of genes that have been faithful to their genomes. Phylogenetic reconstruction is possible despite the massive amounts of horizontal gene transfer

The 3 domains of life



References

- Avise, J. C. 1989. Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution* **43**:1192-1208.
- Bremer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution* **42**:795-803.
- Cavalli-Sforza, L. L., and A. W. Edwards. 1967. Phylogenetic analysis. Models and estimation procedures. *Am J Hum Genet* **19**:233-257.
- Eck, R. V., and M. O. Dayhoff. 1966. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science* **152**:363-366.
- Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics* **25**:471-492.
- Felsenstein, J. 1978. The number of evolutionary trees. *Systematic Zoology* **27**:27-33.
- Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**:368-376.
- Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**:783-791.
- Fitch, W. M. 1971. Toward defining the course of evolution: minimum change for a specified tree topology. *Systematic Zoology* **20**:406-416.
- Hall, B. G. 2001. *Phylogenetic Trees Made Easy: A How-to Manual for Molecular Biologists*. Sinauer Associates, Inc, Sunderland, MA.
- Holder, M. T., and P. O. Lewis. 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nature Reviews Genetics* **4**:275-284.
- Hudson, R. 1983. Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37**:203-217.
- Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**:754-755.
- Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**:2310-2314.
- Li, W.-H. 1997. *Molecular Evolution*. Sinauer, Sunderland, MA.
- Nichols, R. 2001. Gene trees and species trees are not the same. *Trends in Ecology and Evolution* **16**:358-364.
- Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5**:568-583.

- Rannala, B., and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol* **43**:304-311.
- Rzhetsky, A., and M. Nei. 1992. A simple method for estimating and testing minimum-evolution trees. *Mol Biol Evol* **9**:945-967.
- Saitou, N., and M. Nei. 1987. The neighbor-joining method : A new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**:406-425.
- Sneath, P. H. A., and R. R. Sokal. 1973. *Numerical Taxonomy*. W. H. Freeman, San Francisco, CA.
- Yang, Z., and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo Method. *Molecular Biology and Evolution* **14**:717-724.