

CAP 5510: Introduction to Bioinformatics
CGS 5166: Bioinformatics Tools

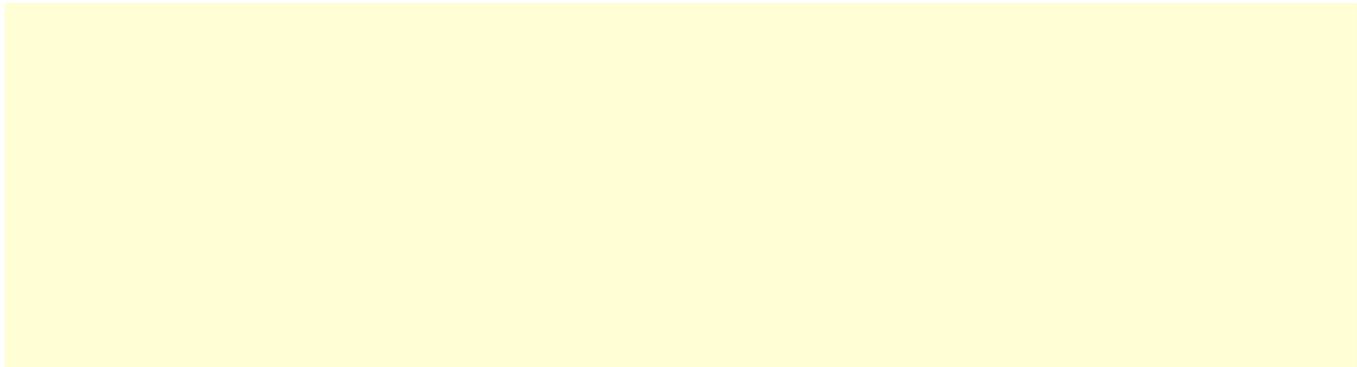
Giri Narasimhan

ECS 254; Phone: x3748

giri@cis.fiu.edu

www.cis.fiu.edu/~giri/teach/BioinfS15.html

Evolution and Phylogeny



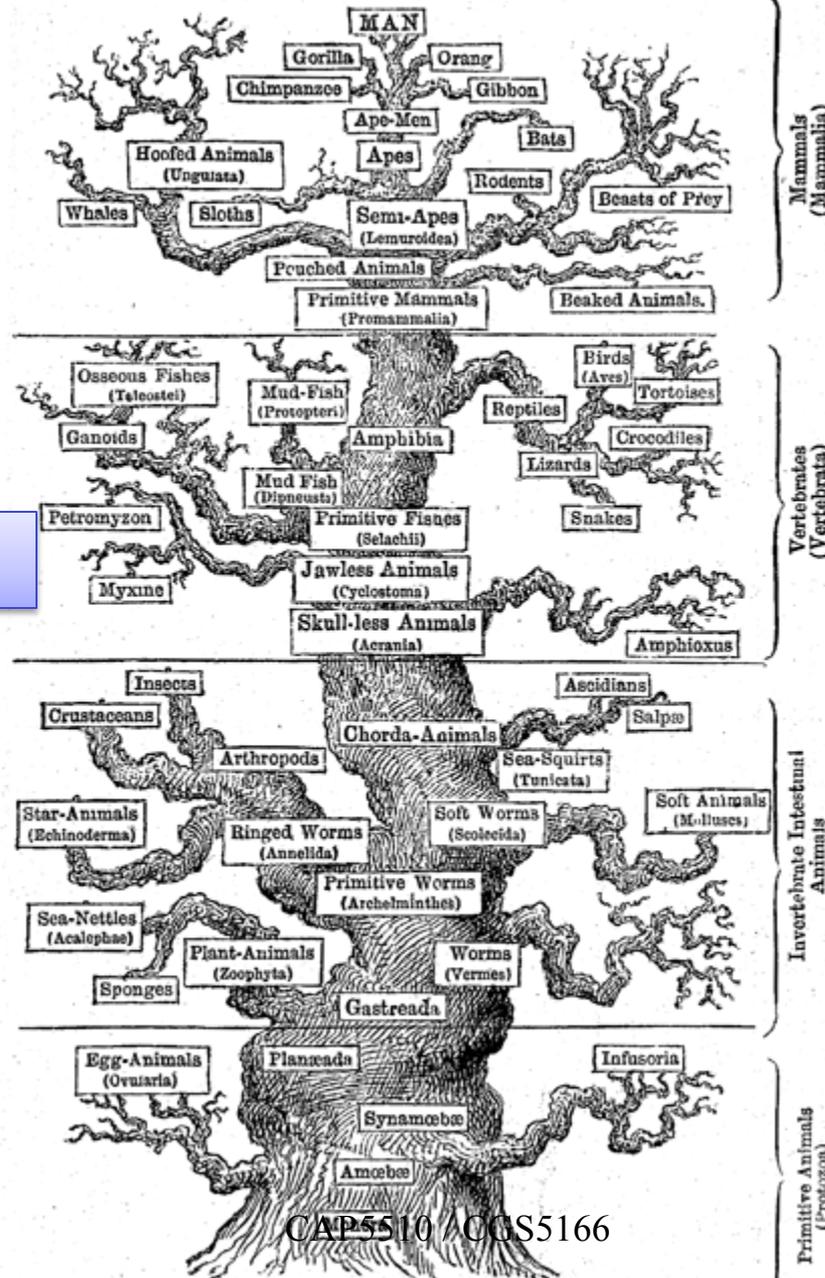
Darwin: Evolution & Natural Selection

- ❑ Charles Darwin's 1859 book (*On the Origin of Species By Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*) introduced the **Theory of Evolution**.
- ❑ Struggle for existence induces a natural selection. Offspring are dissimilar from their parents (that is, variability exists), and individuals that are more fit for a given environment are selected for. In this way, over long periods of time, species evolve. Groups of organisms change over time so that descendants differ structurally and functionally from their ancestors.

Dominant View of Evolution

- All existing organisms are derived from a common ancestor and that new species arise by splitting of a population into subpopulations that do not cross-breed.
- Organization: **Directed Rooted Tree**; Existing species: **Leaves**; Common ancestor species (divergence event): **Internal node**; Length of an edge: **Time**.

PEDIGREE OF MAN.



Five kingdom system (Haeckel, 1879)

Slide by Pevsner

mammals

vertebrates

invertebrates

protozoa

- animals
- plants
- fungi
- protists
- monera

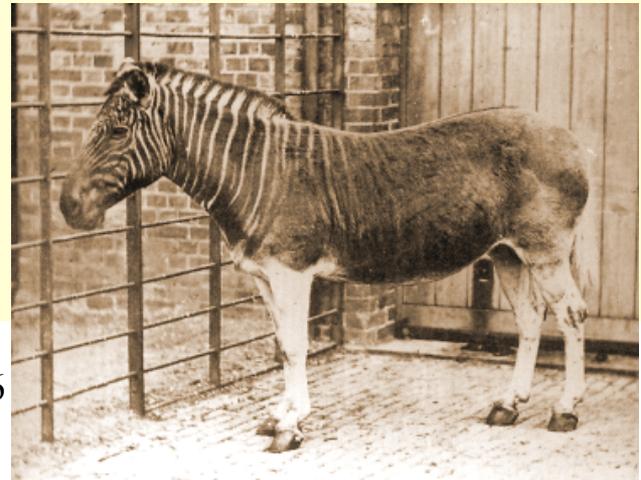
Evolution & Phylogeny

- ❑ At the molecular level, evolution is a process of mutation with selection.
- ❑ Molecular evolution is the study of changes in genes and proteins throughout different branches of the tree of life.
- ❑ Phylogeny is the inference of evolutionary relationships. Traditionally, phylogeny relied on the comparison of morphological features between organisms. Today, molecular sequence data are also used for phylogenetic analyses.

Questions for Phylogenetic Analysis

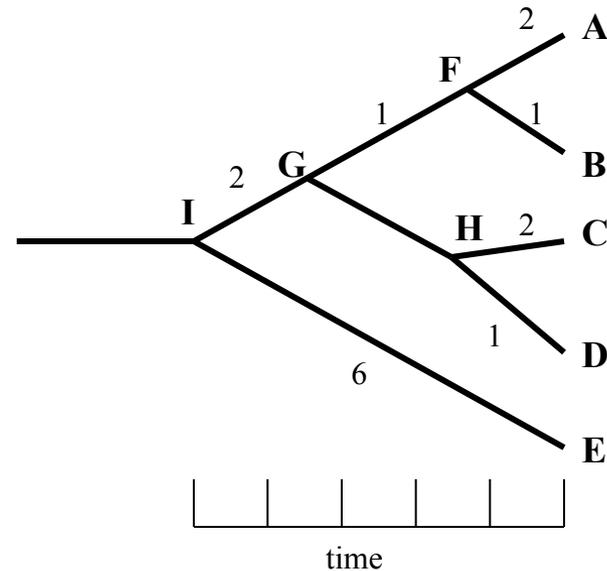
- How many genes are related to my favorite gene?
- How related are whales, dolphins & porpoises to cows?
- Where and when did HIV or other viruses originate?
- What is the history of life on earth?
- Was the extinct quagga more like a zebra or a horse?

Slide by Pevsner



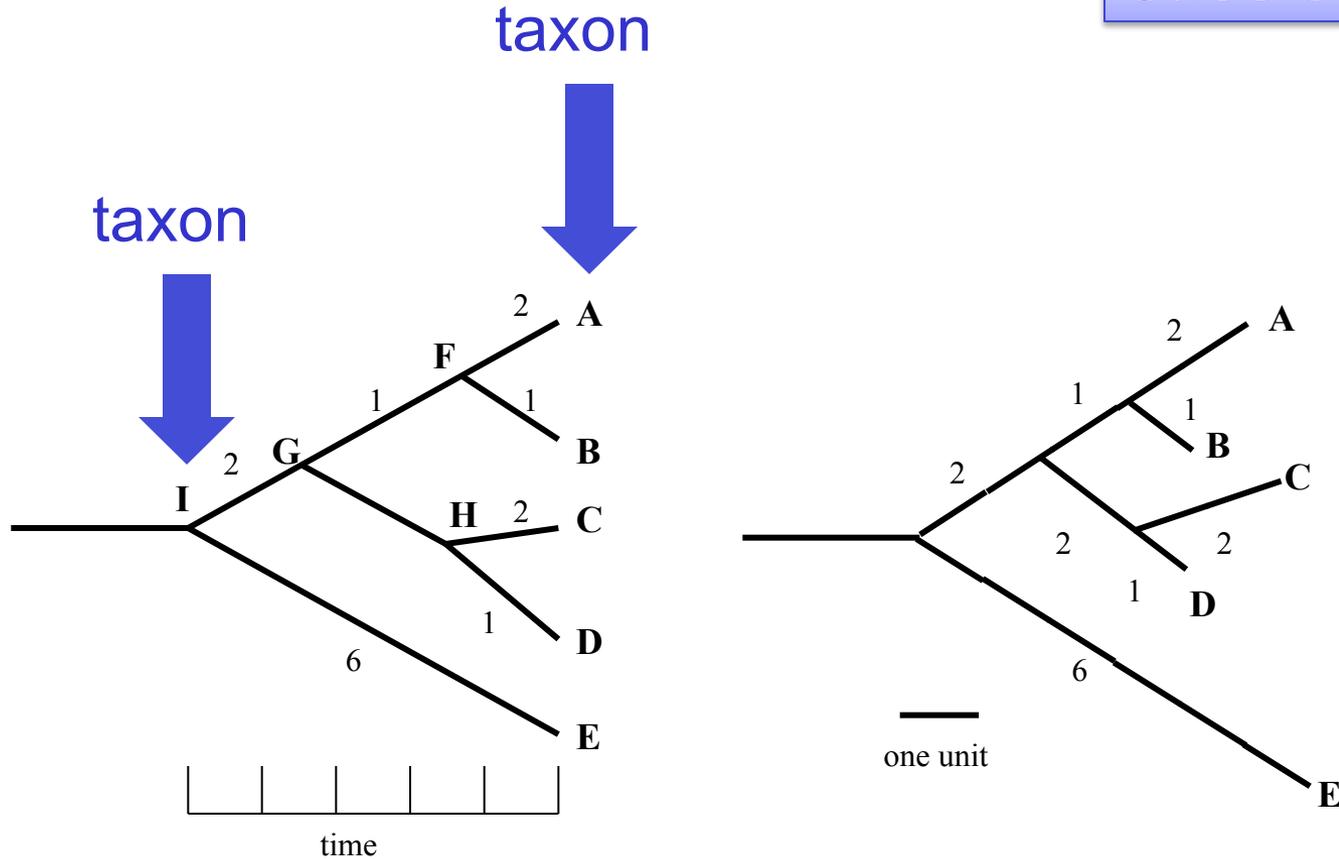
Phylogenetic Trees

□ Molecular phylogeny uses trees to depict evolutionary relationships among organisms. These trees are based upon DNA and protein sequence data.



Tree Nomenclature

Slide by Pevsner



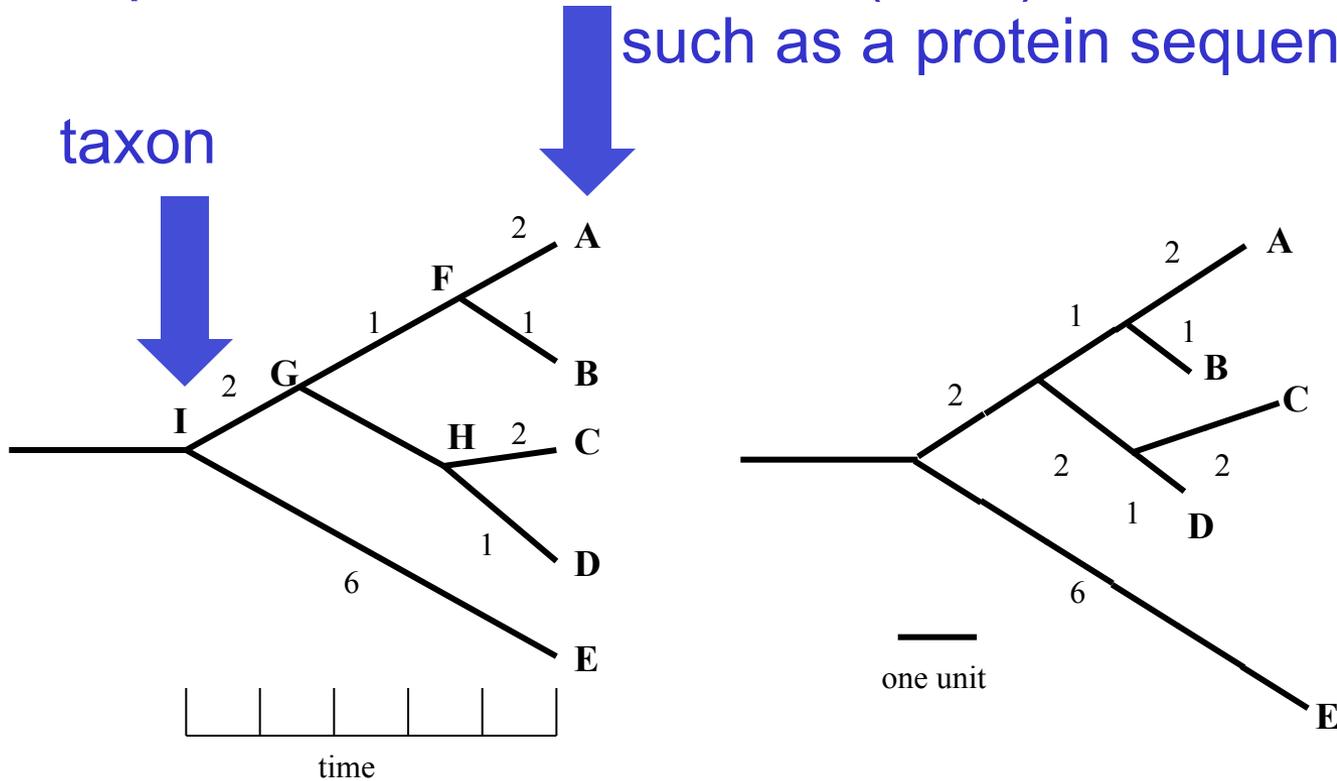
Tree nomenclature

Slide by Pevsner

operational taxonomic unit (OTU)

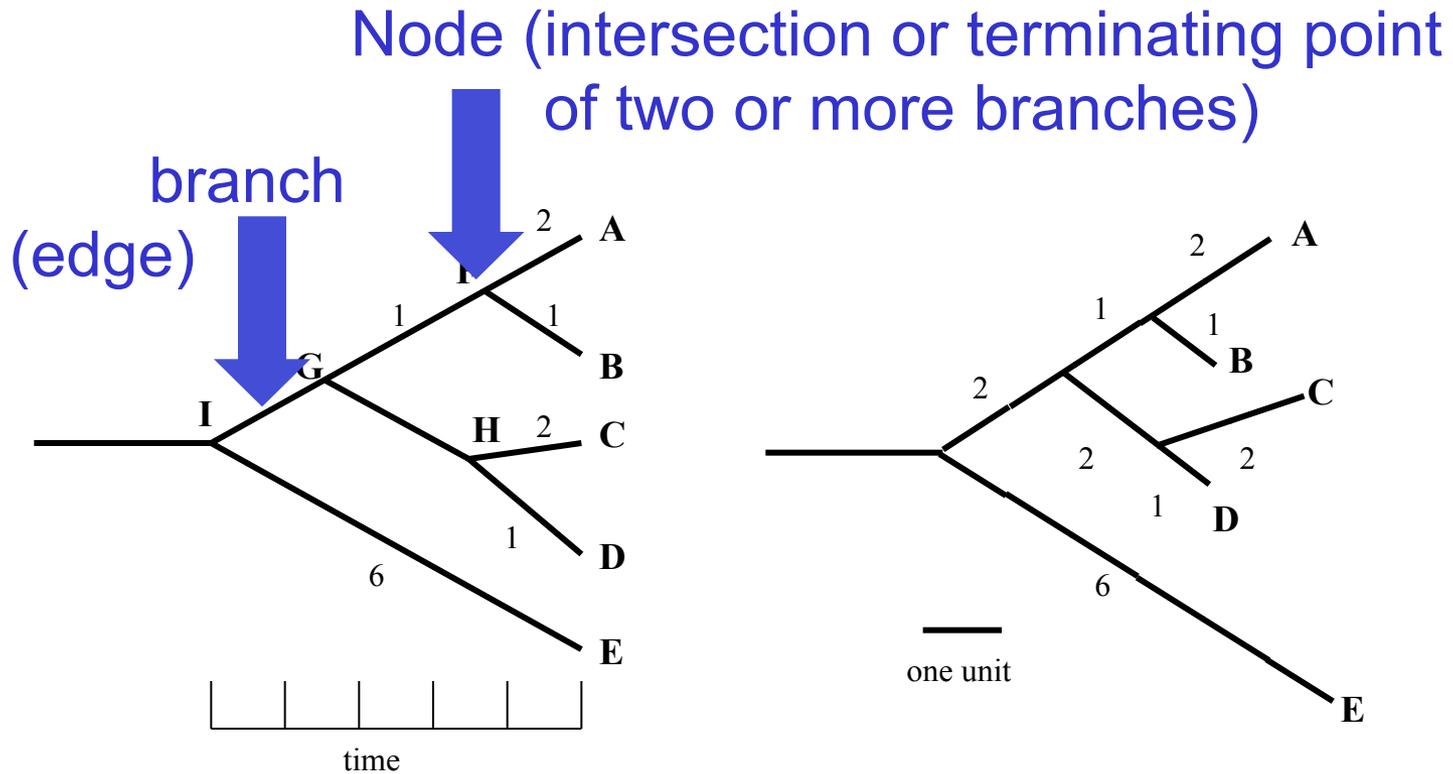
such as a protein sequence

taxon



Tree nomenclature

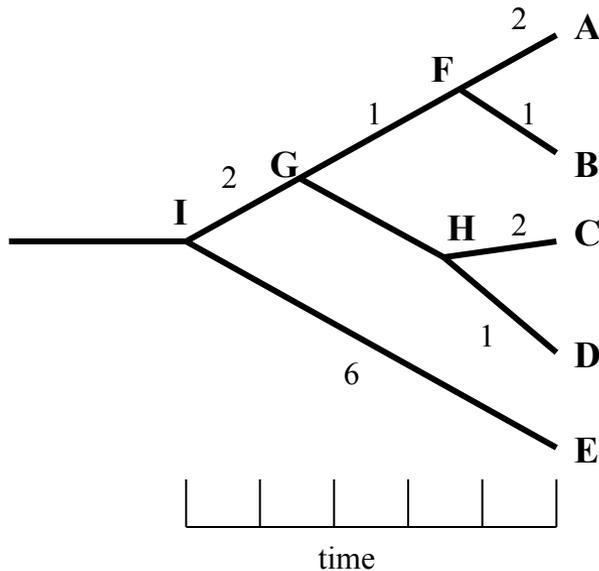
Slide by Pevsner



Tree nomenclature

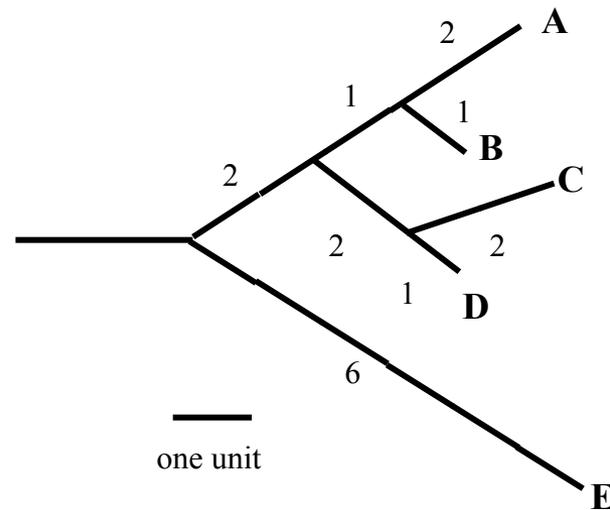
Slide by Pevsner

Branches are unscaled...



...OTUs are neatly aligned,
and nodes reflect time

Branches are scaled...

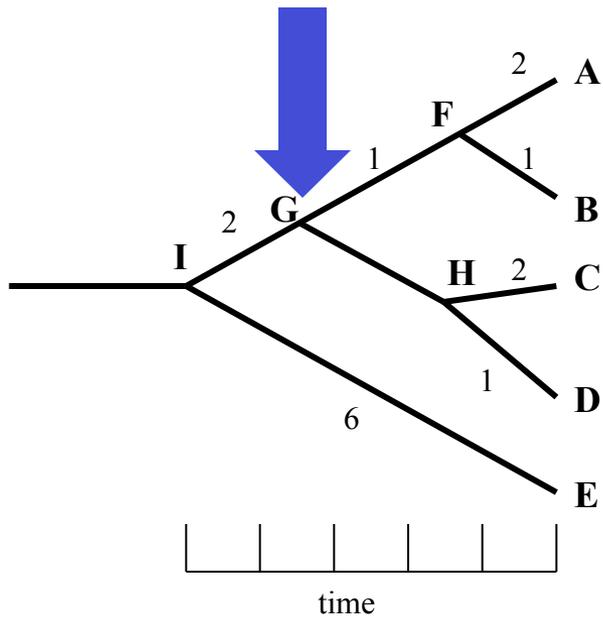


...branch lengths are
proportional to number of
amino acid changes

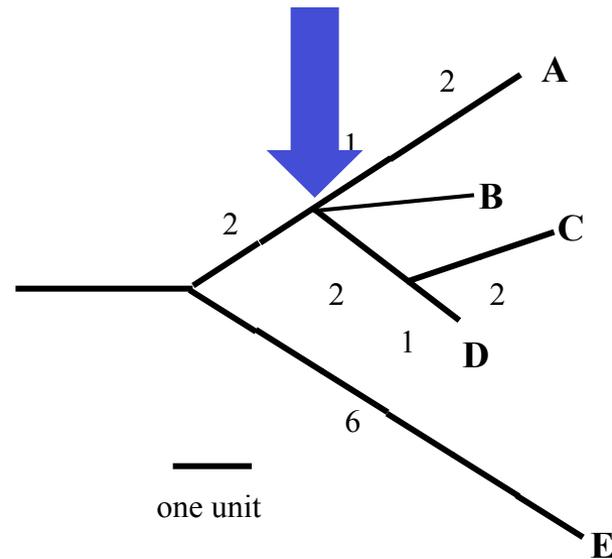
Tree nomenclature

Slide by Pevsner

bifurcating
internal
node

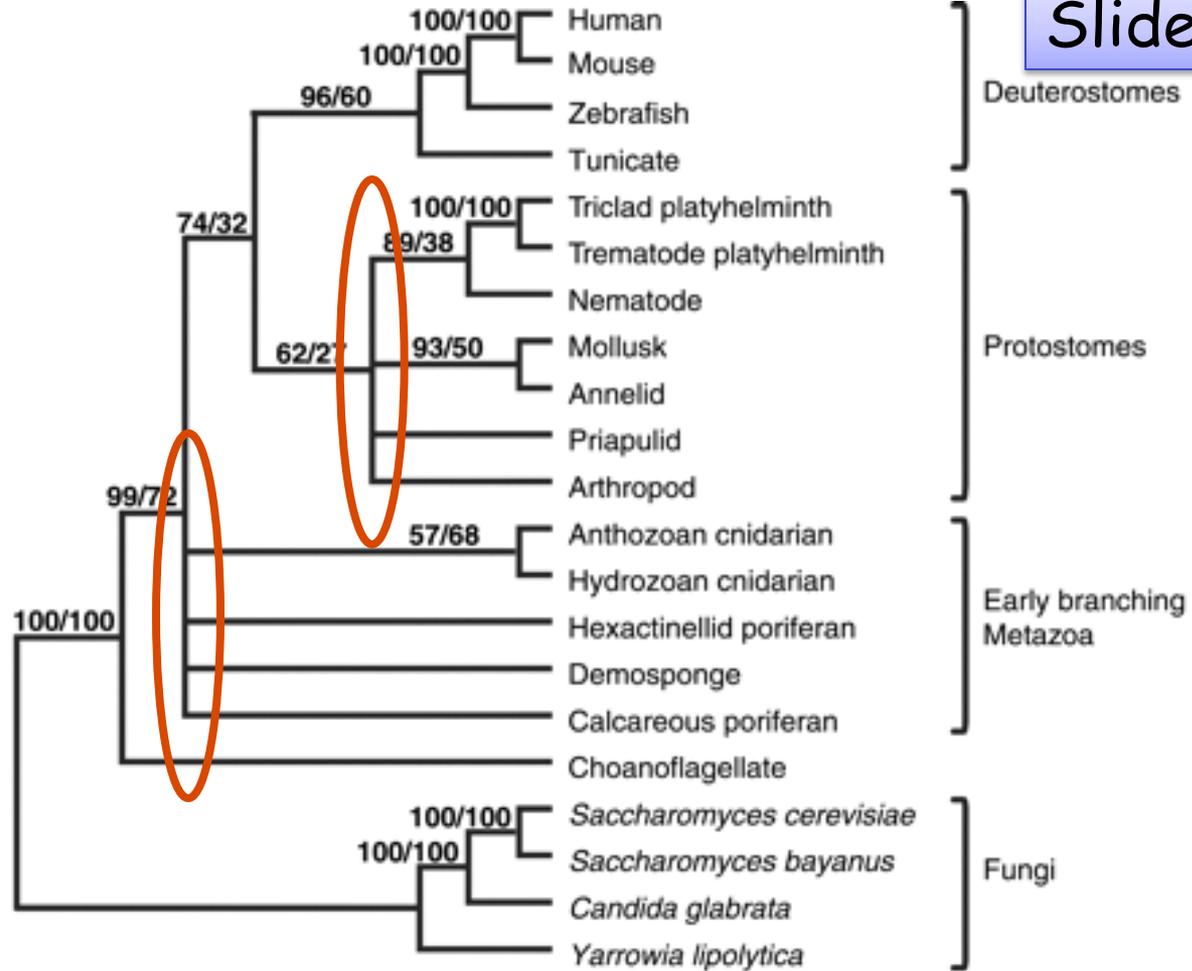


multifurcating
internal
node



Examples of multifurcation: failure to resolve the branching order of some metazoans and protostomes

Slide by Pevsner

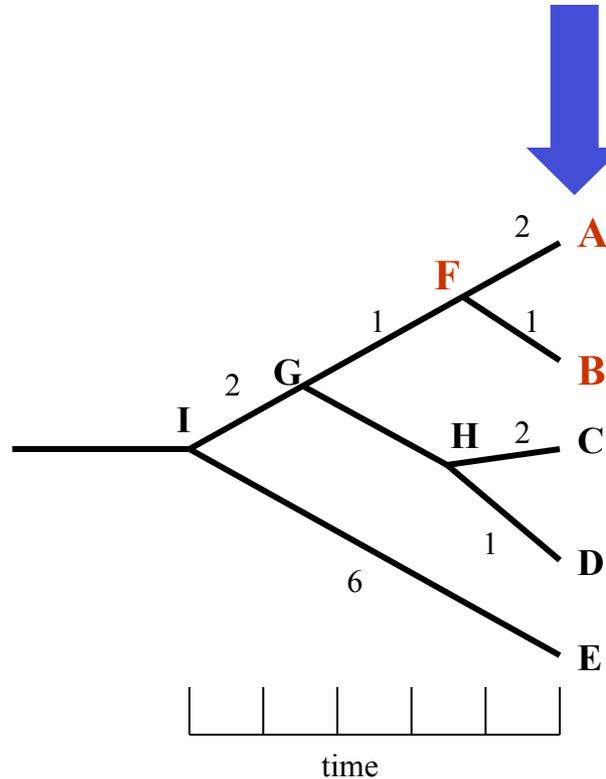


Rokas A. et al., Animal Evolution and the Molecular Signature of Radiations Compressed in Time, *Science* 310:1933 (2005), Fig. 1.

Tree nomenclature: clades

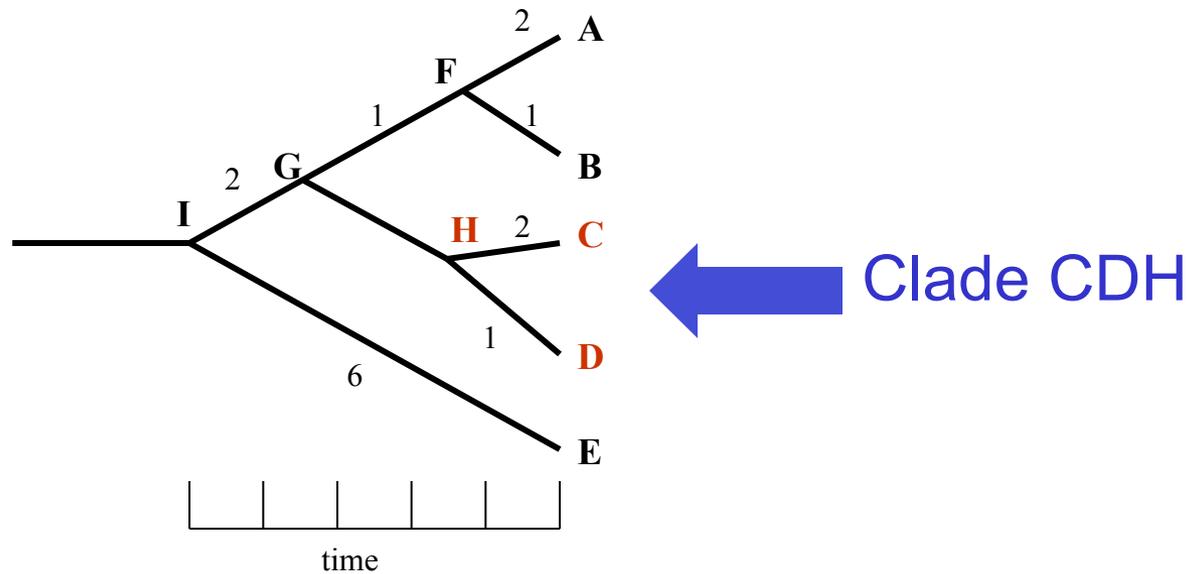
Slide by Pevsner

Clade ABF (monophyletic group)



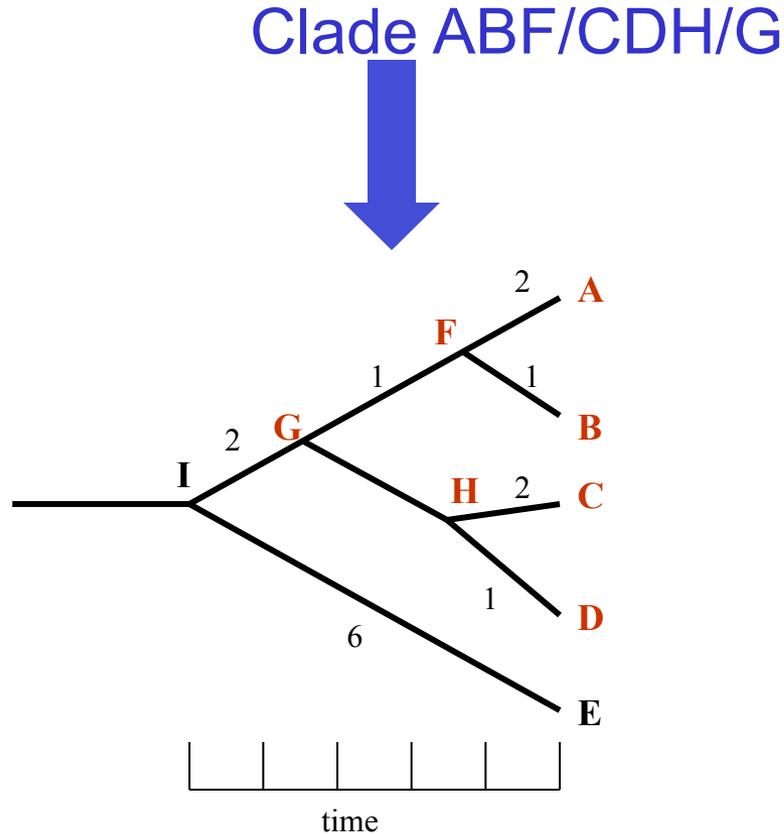
Tree nomenclature

Slide by Pevsner



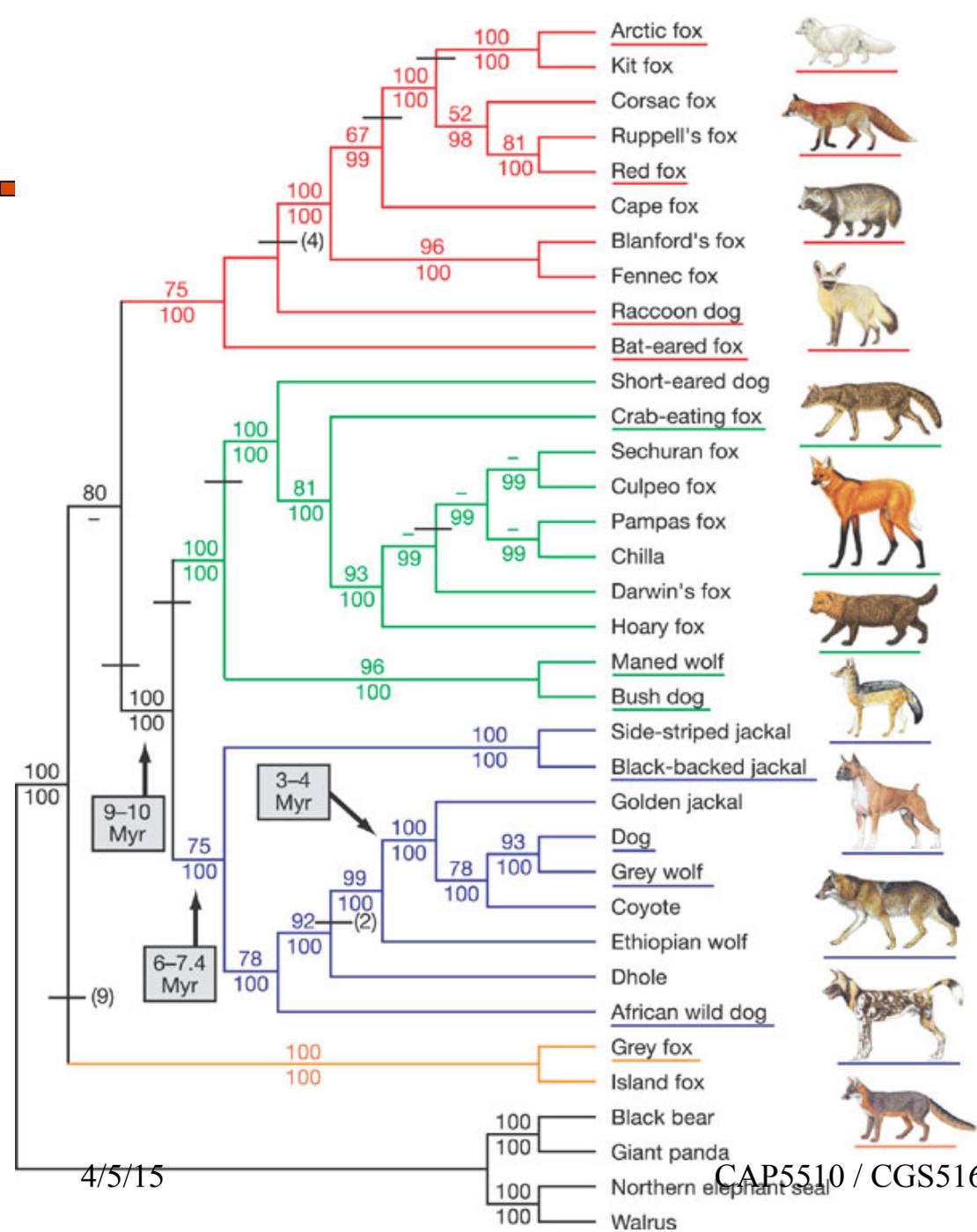
Tree nomenclature

Slide by Pevsner



Examples of clades

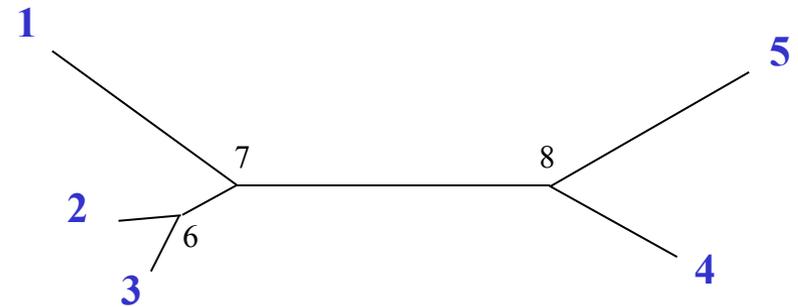
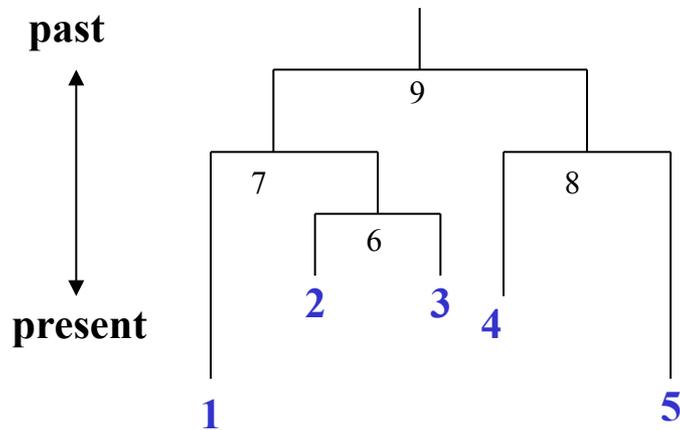
Slide by Pevsner



Lindblad-Toh et al., *Nature* 438: 803 (2005), fig. 10

Tree nomenclature: roots

Slide by Pevsner

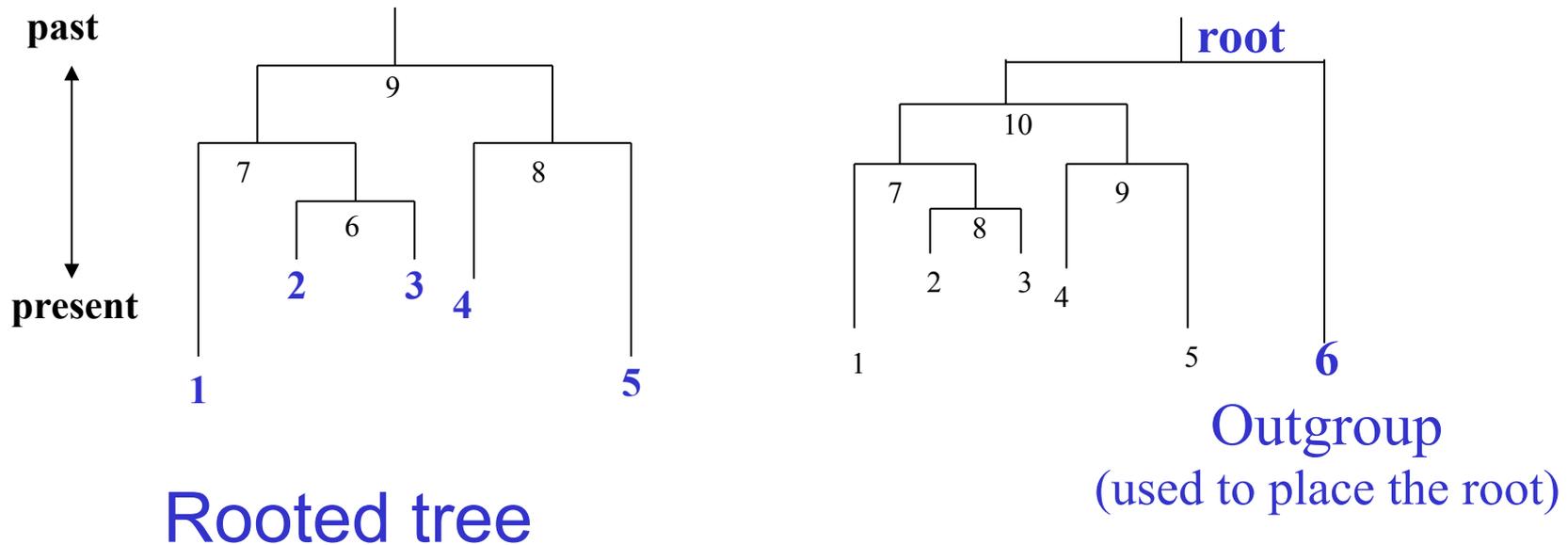


Rooted tree
(specifies evolutionary
path)

Unrooted tree

Tree nomenclature: outgroup rooting

Slide by Pevsner



□ 2 broad categories:

● Distance-based methods

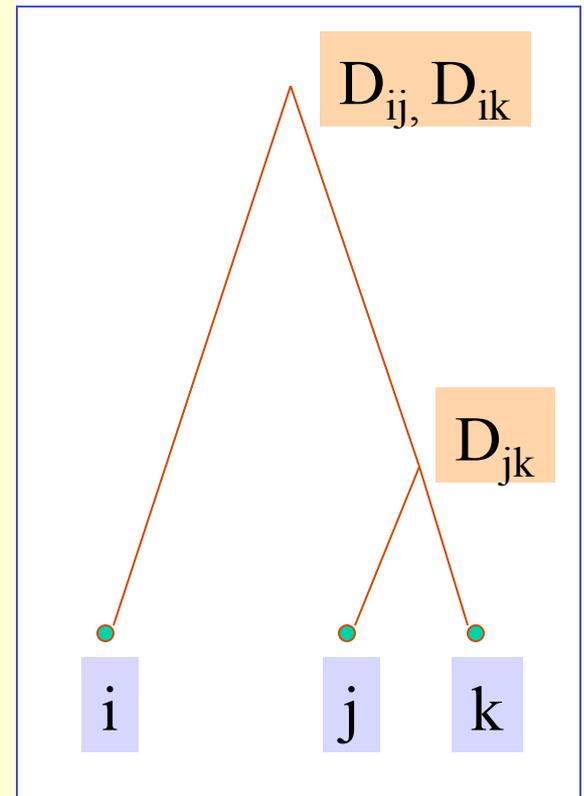
- Ultrametric
- Additive:
 - UPGMA
 - Transformed Distance
 - Neighbor-Joining

● Character-based

- Maximum Parsimony
- Maximum Likelihood
- Bayesian Methods

Ultrametric

- An ultrametric tree:
 - decreasing internal node labels
 - distance between two nodes is label of least common ancestor.
- An ultrametric distance matrix:
 - Symmetric matrix such that for every i, j, k , there is **tie for maximum** of $D(i,j), D(j,k), D(i,k)$



Ultrametric: Assumptions

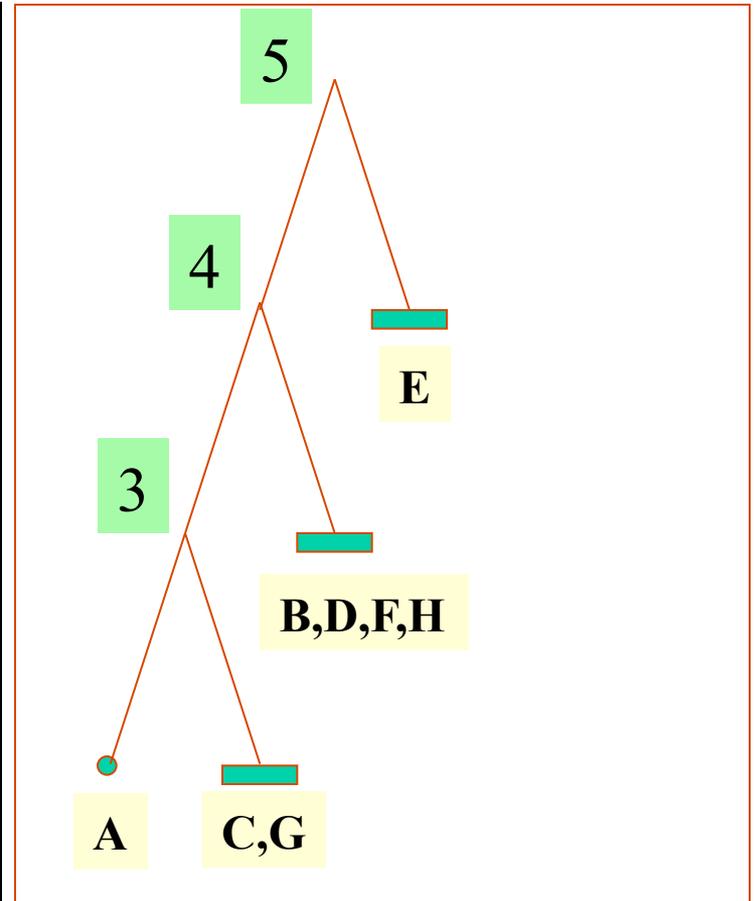
- **Molecular Clock Hypothesis**, Zuckerkandl & Pauling, 1962: Accepted point mutations in amino acid sequence of a protein occurs at a **constant** rate.
 - Varies from protein to protein
 - Varies from one part of a protein to another

Ultrametric Data Sources

- Lab-based methods: **hybridization**
 - Take denatured DNA of the 2 taxa and let them hybridize. Then measure energy to separate.
- Sequence-based methods: **distance**

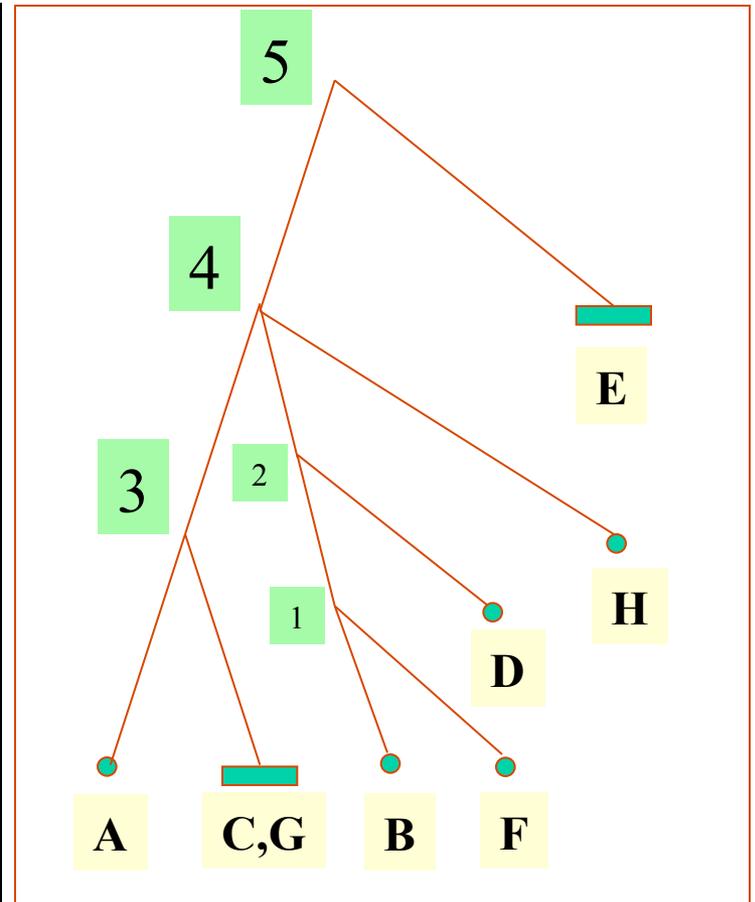
Ultrametric: Example

	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B								
C								
D								
E								
F								
G								
H								



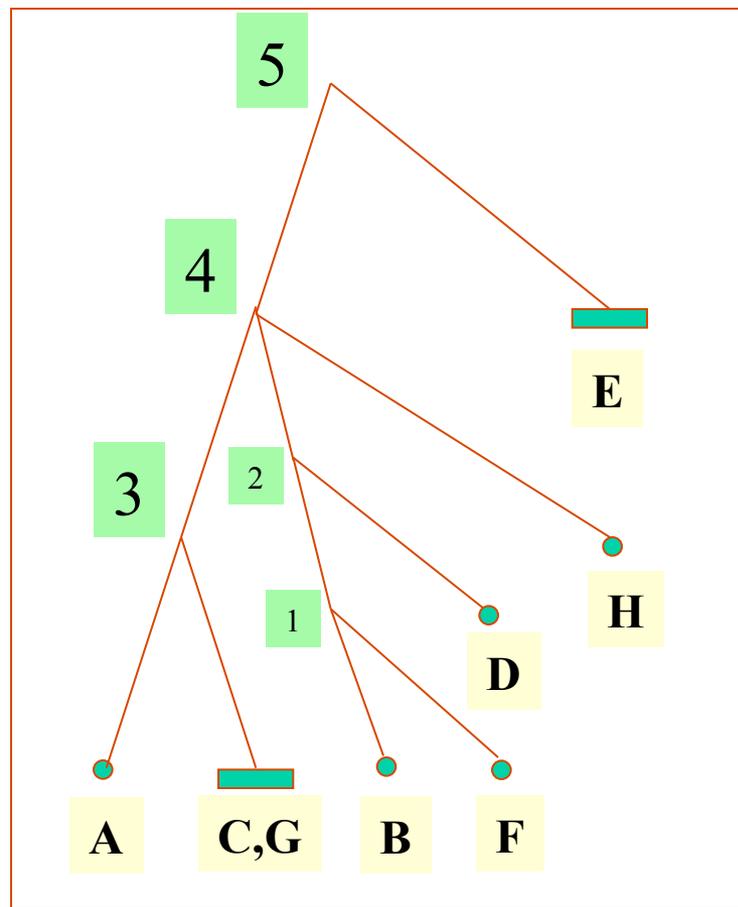
Ultrametric: Example

	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B		0	4	2	5	1	4	4
C								
D								
E								
F								
G								
H								



Ultrametric: Distances Computed

	A	B	C	D	E	F	G	H
A	0	4	3	4	5	4	3	4
B		0	4	2	5	1	4	4
C							2	
D								
E								
F								
G								
H								



Ultrametric: Assumptions

- **Molecular Clock Hypothesis**, Zuckerkandl & Pauling, 1962: Accepted point mutations in amino acid sequence of a protein occurs at a **constant** rate.
 - Varies from protein to protein
 - Varies from one part of a protein to another

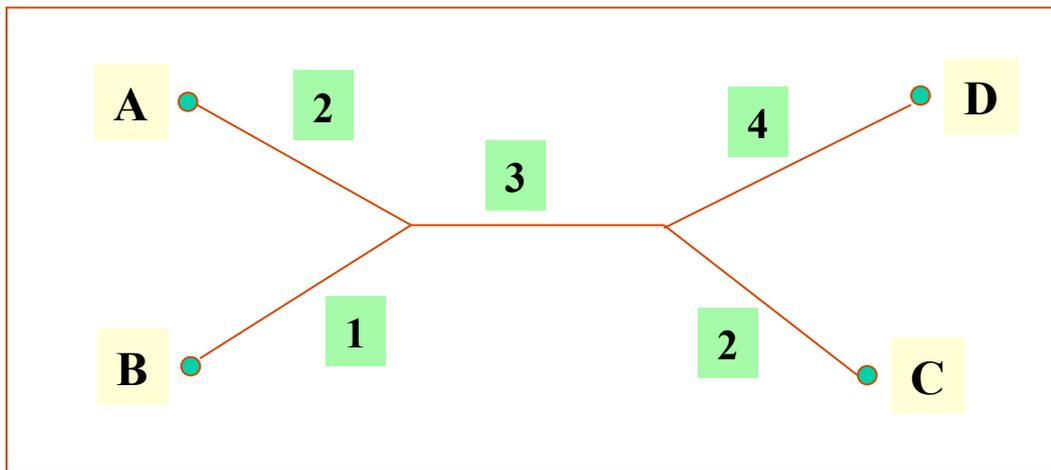
Ultrametric Data Sources

- Lab-based methods: **hybridization**
 - Take denatured DNA of the 2 taxa and let them hybridize. Then measure energy to separate.
- Sequence-based methods: **distance**

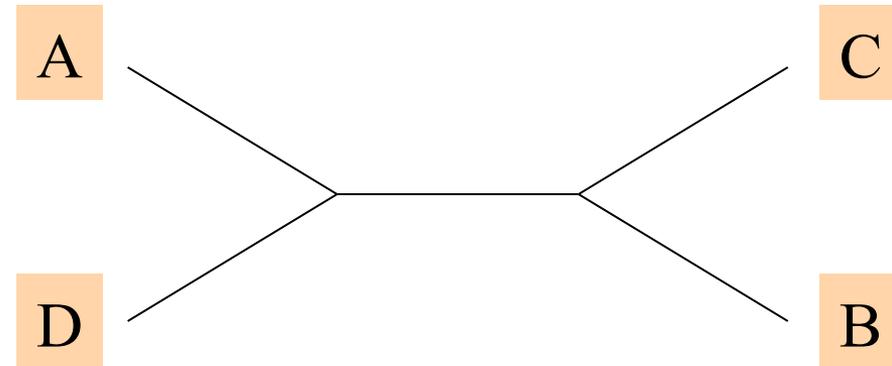
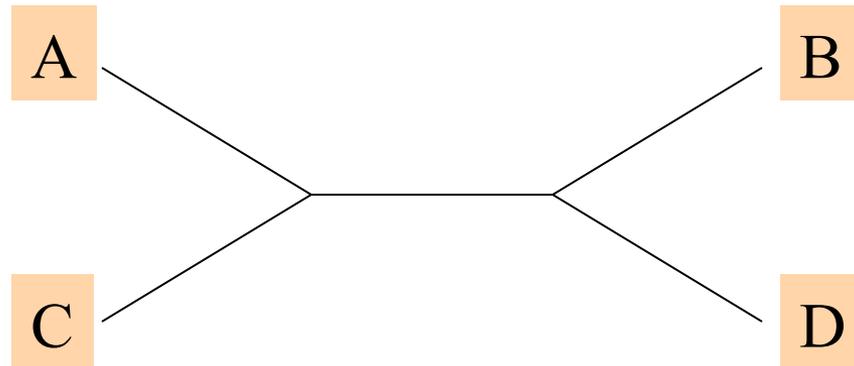
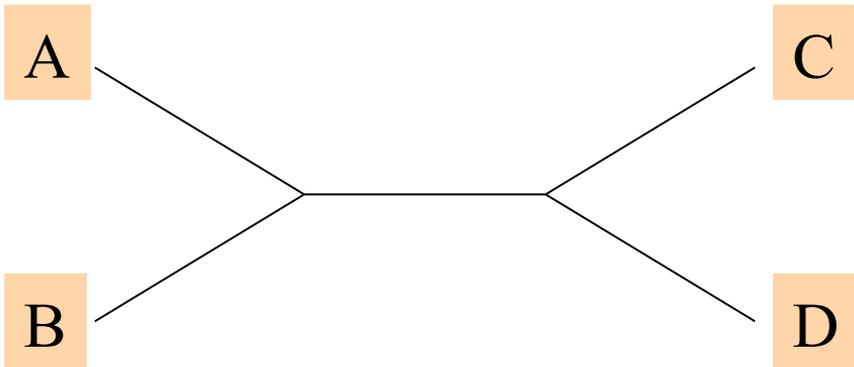
Additive-Distance Trees

Additive distance trees are edge-weighted trees, with distance between leaf nodes are exactly equal to length of path between nodes.

	A	B	C	D
A	0	3	7	9
B		0	6	8
C			0	6
D				0



Unrooted Trees on 4 Taxa

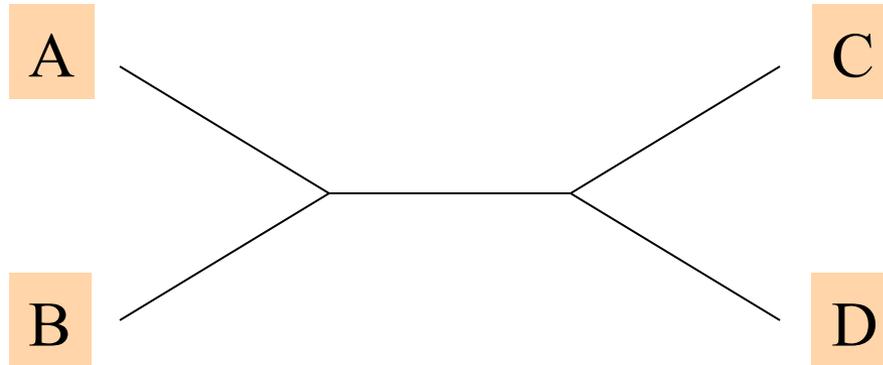


Four-Point Condition

□ If the true tree is as shown below, then

1. $d_{AB} + d_{CD} < d_{AC} + d_{BD}$, and

2. $d_{AB} + d_{CD} < d_{AD} + d_{BC}$

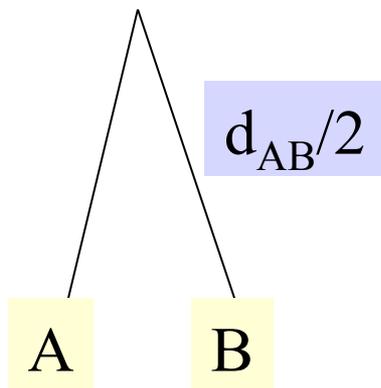


Unweighted pair-group method with arithmetic means (UPGMA)

	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

	AB	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

$$d_{(AB)C} = (d_{AC} + d_{BC}) / 2$$



Transformed Distance Method

- UPGMA makes errors when rate constancy among lineages does not hold.
- Remedy: introduce an outgroup & make corrections

$$D_{ij}' = \frac{D_{ij} - D_{iO} - D_{jO}}{2} + \left(\frac{\sum_{k=1}^n D_{kO}}{n} \right)$$

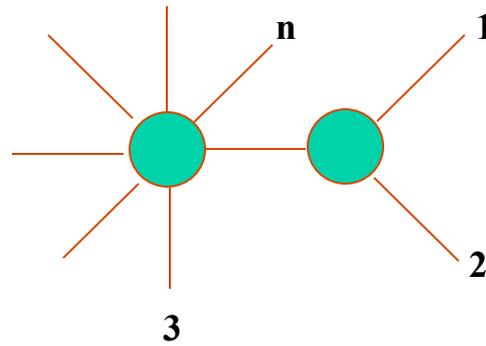
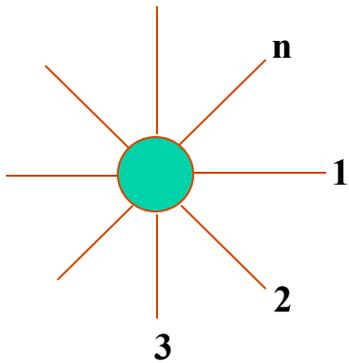
- Now apply UPGMA

Saitou & Nei: Neighbor-Joining Method

- Start with a **star topology**.
- Find the pair to separate such that the total length of the tree is minimized. The pair is then replaced by its arithmetic mean, and the process is repeated.

$$S_{12} = \frac{D_{12}}{2} + \frac{1}{2(n-2)} \sum_{k=3}^n (D_{1k} + D_{2k}) + \frac{1}{(n-2)} \sum_{3 \leq i \leq j \leq n} D_{ij}$$

Neighbor-Joining



$$S_{12} = \frac{D_{12}}{2} + \frac{1}{2(n-2)} \sum_{k=3}^n (D_{1k} + D_{2k}) + \frac{1}{(n-2)} \sum_{3 \leq i \leq j \leq n} D_{ij}$$

Constructing Evolutionary/Phylogenetic Trees

□ 2 broad categories:

● Distance-based methods

- Ultrametric
- Additive:
 - UPGMA
 - Transformed Distance
 - Neighbor-Joining

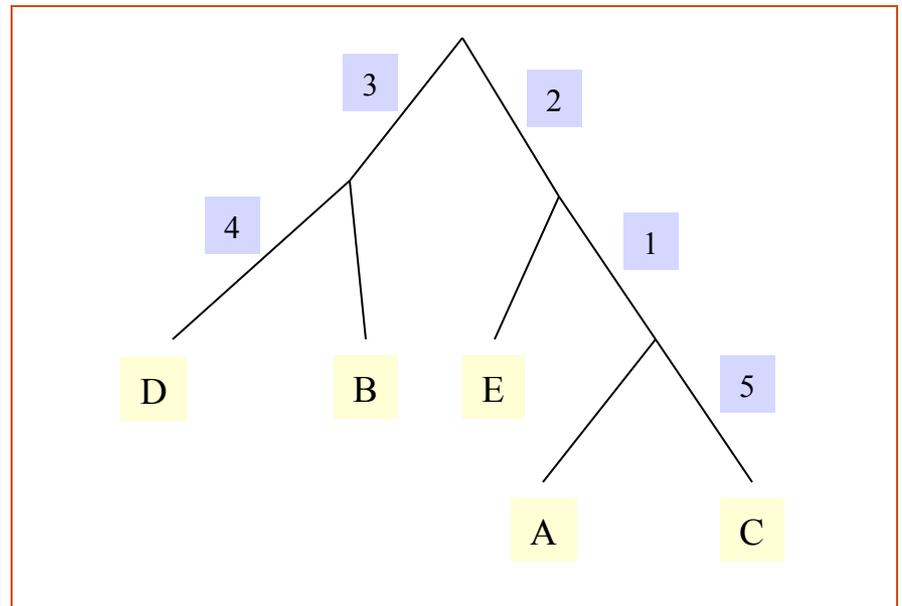
● Character-based

- Maximum Parsimony
- Maximum Likelihood
- Bayesian Methods

Character-based Methods

- ❑ Input: characters, morphological features, sequences, etc.
- ❑ Output: phylogenetic tree that provides the history of what features changed. [Perfect Phylogeny Problem]
- ❑ one leaf/object, 1 edge per character, path \Leftrightarrow changed traits

	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	0
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	0

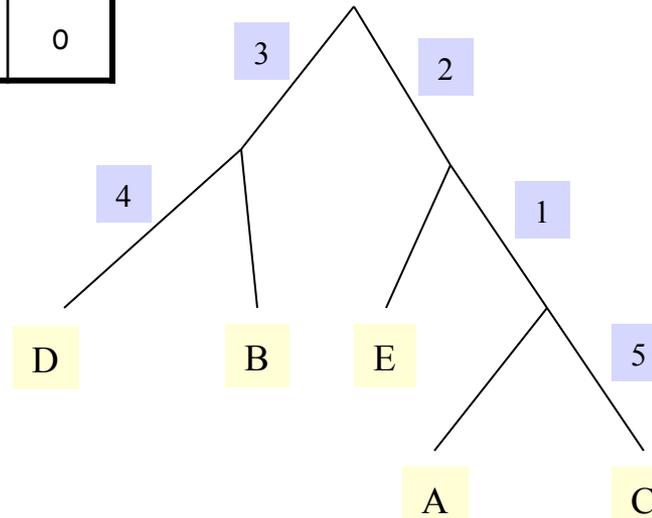


Example

❑ Perfect phylogeny does not always exist.

	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	0
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	0

	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	1
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	1



Maximum Parsimony

- Minimize the total number of mutations implied by the evolutionary history

Examples of Character Data

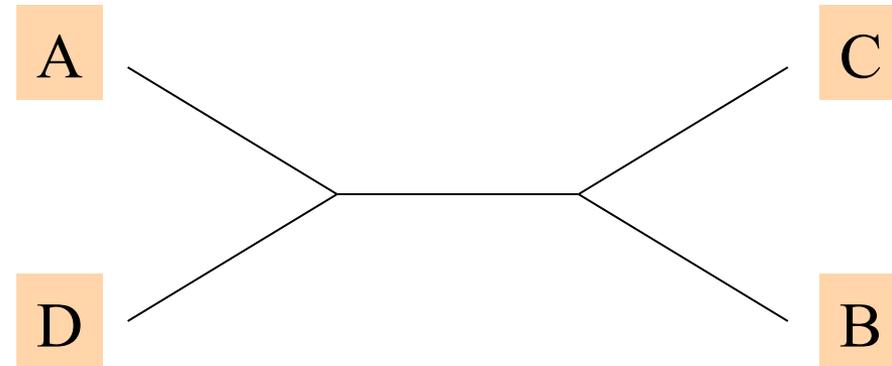
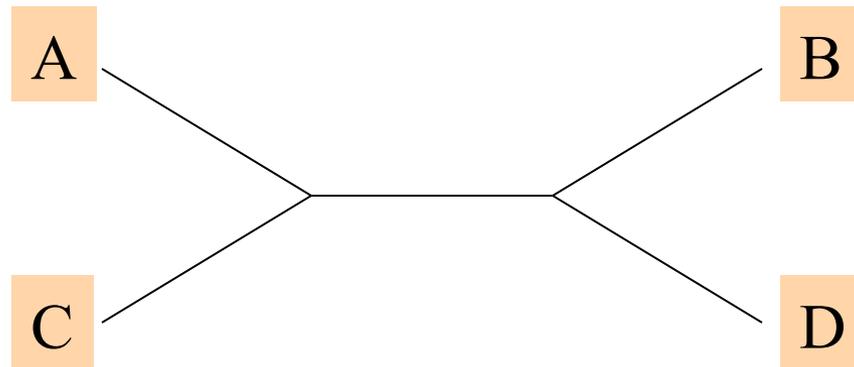
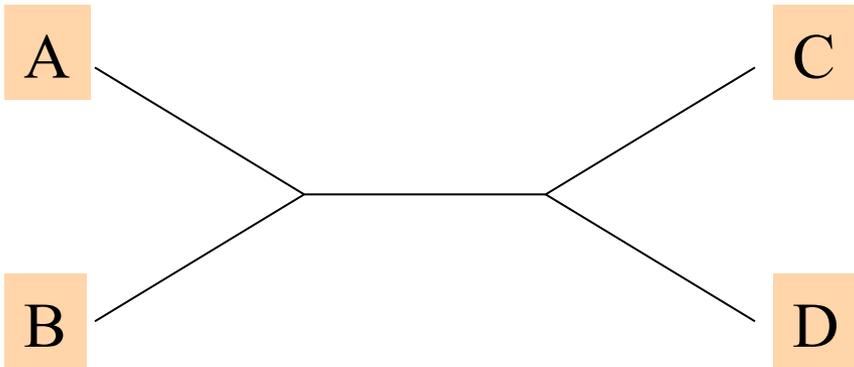
	1	2	3	4	5
A	1	1	0	0	0
B	0	0	1	0	1
C	1	1	0	0	1
D	0	0	1	1	0
E	0	1	0	0	1

	Characters/Sites								
Sequences	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

Maximum Parsimony Method: Example

	Characters/Sites								
Sequences	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	T	C	A
2	A	G	C	C	G	T	T	C	T
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	T

Unrooted Trees on 4 Taxa



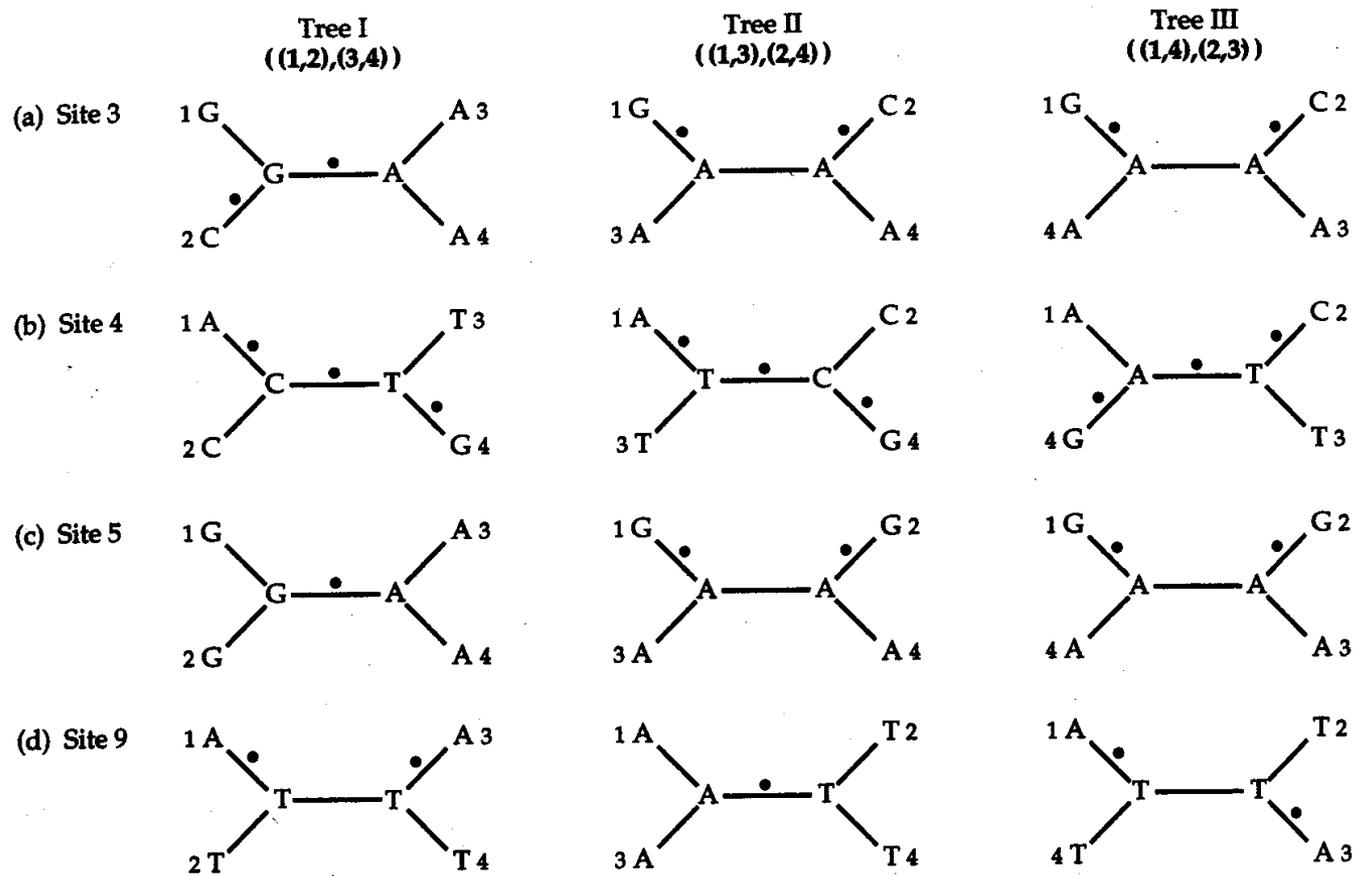


FIGURE 5.14 Three possible unrooted trees (I, II, and III) for four DNA sequences (1, 2, 3, and 4) that have been used to choose the most parsimonious tree. The possible phylogenetic relationships among the four sequences are shown in Newick format. The terminal nodes are marked by the sequence number and the nucleotide type at homologous positions in the extant species. Each dot on a branch means a substitution is inferred on that branch. Note that the nucleotides at the two internal nodes of each tree represent one possible reconstruction from among several alternatives. For example, the nucleotides at both the internal nodes of tree III(d) (bottom right) can be A instead of T. In this case, the two substitutions will be positioned on the branches leading to species 2 and 4. Alternatively, other combinations of nucleotides can be placed at the internal nodes. However, these alternatives will require three substitutions or more. The minimum number of substitutions required for site 9 is two.

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

Inferring nucleotides on internal nodes

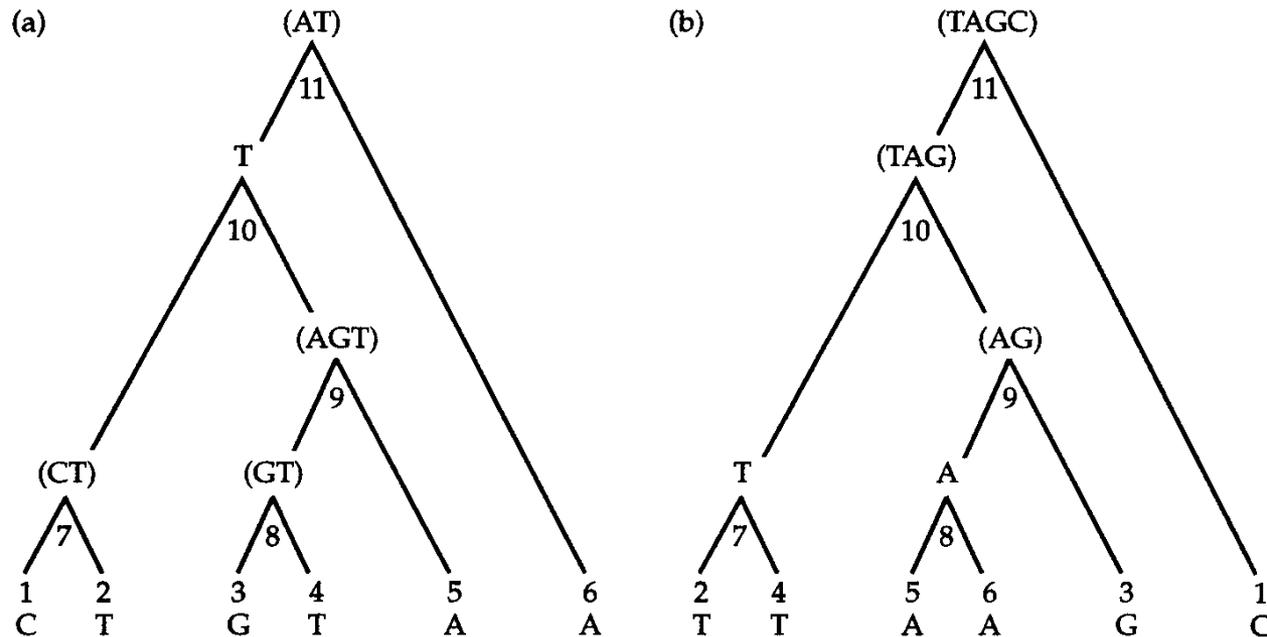


FIGURE 5.15 Nucleotides in six extant species (1–6) and inferred possible nucleotides in five ancestral species (7–11) according to the method of Fitch (1971). Unions are indicated by parentheses. Two different trees (a and b) are depicted. Note that the inference of an ancestral nucleotide at an internal node is dependent on the tree. Modified from Fitch (1971).

Searching for the Maximum Parsimony Tree: Exhaustive Search

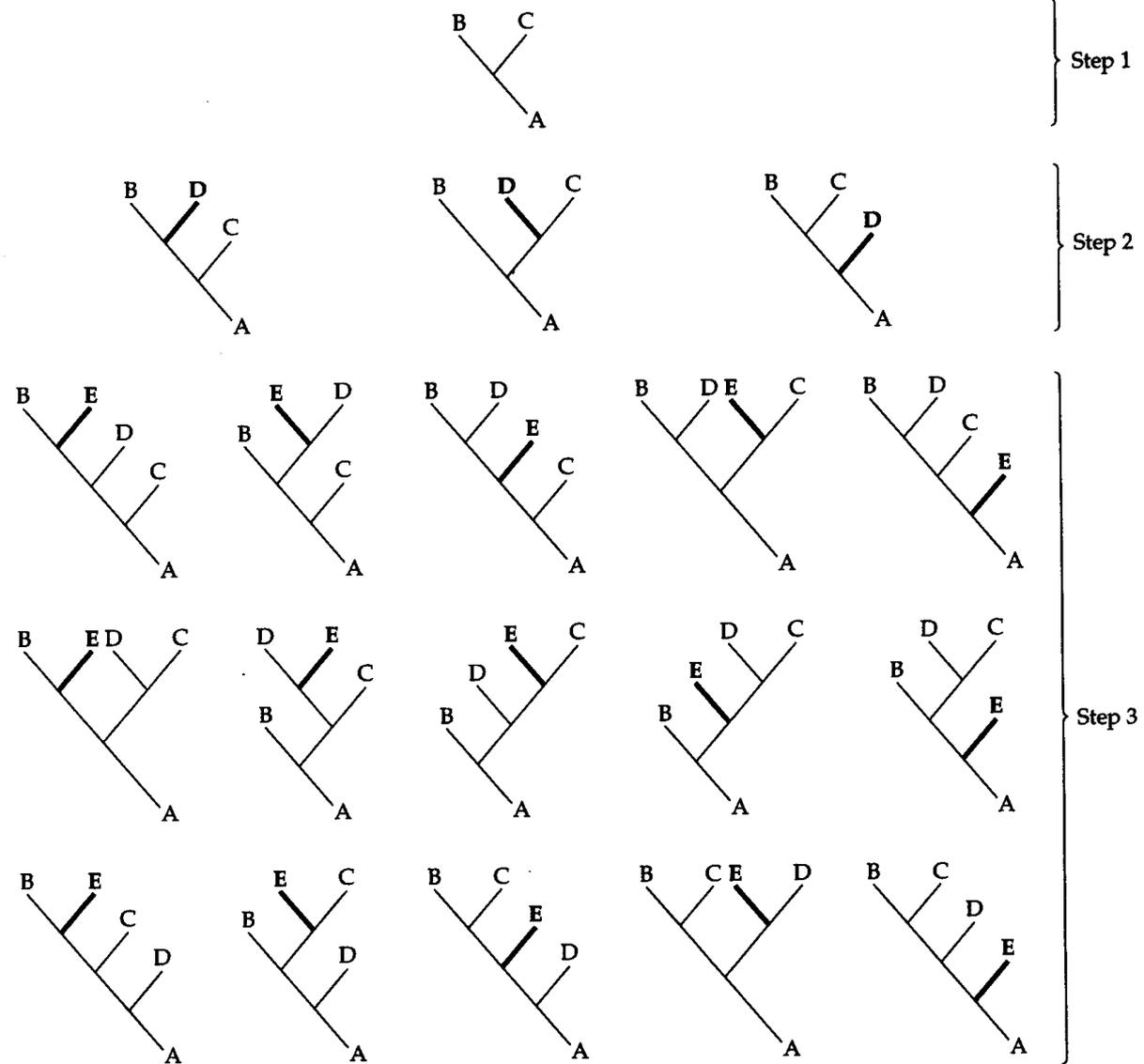
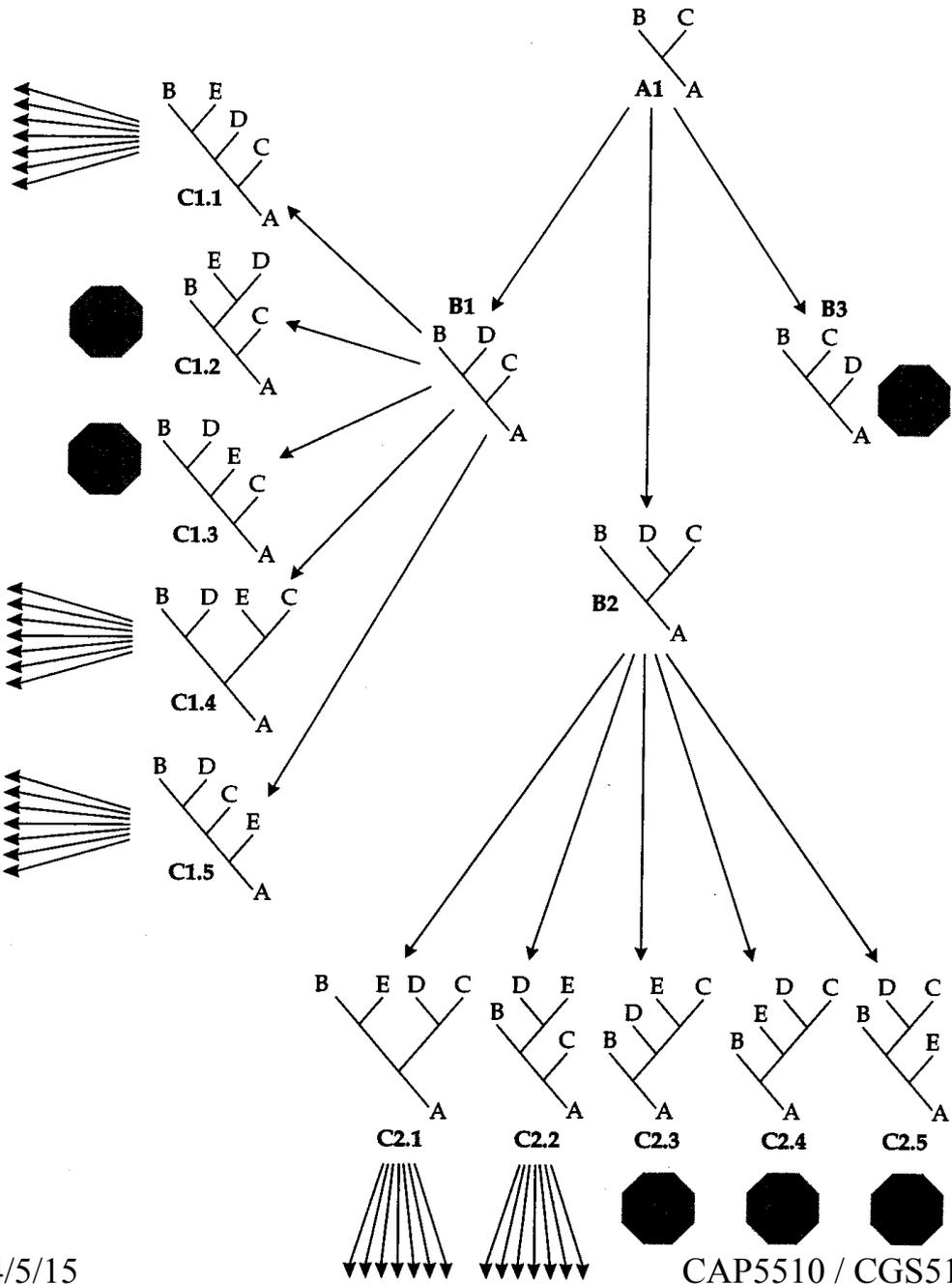


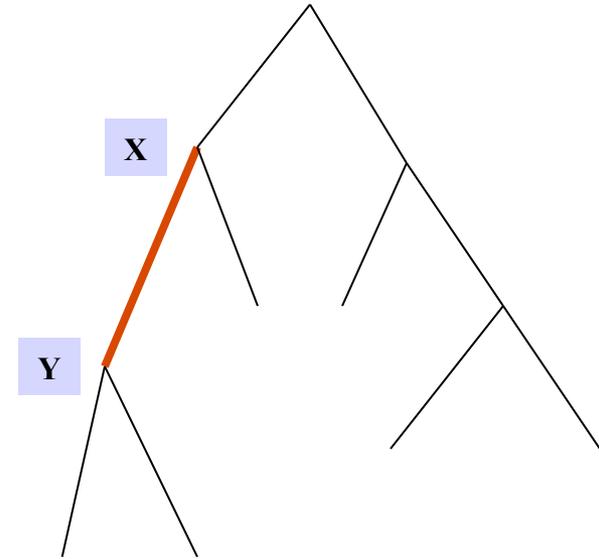
FIGURE 5.16 Exhaustive stepwise construction of all 15 possible trees for five OTUs. In step 1, we form the only possible unrooted tree for the first three OTUs (A, B, and C). In step 2, we add OTU D to each of the three branches of the tree in step 1, thereby generating three unrooted trees for four OTUs. In step 3, we add OTU E to each of the five branches of the three trees in step 2, thereby generating 15 unrooted trees. Additions of OTUs are shown as heavier lines. Modified from Swofford et al. (1996).



Searching for the Maximum Parsimony Tree:
Branch-&-Bound

Probabilistic Models of Evolution

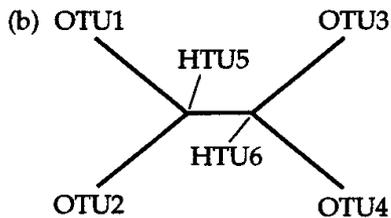
- Assuming a **model of substitution**,
 - $\Pr\{S_i(t+\Delta) = Y \mid S_i(t) = X\}$,
- Using this formula it is possible to compute the likelihood that data D is generated by a given phylogenetic tree T under a model of substitution. Now find the tree with the maximum likelihood.



- Time elapsed? Δ
- Prob of change along edge?
 $\Pr\{S_i(t+\Delta) = Y \mid S_i(t) = X\}$
- Prob of data? **Product of prob for all edges**

FIGURE 5.19 Schematic representation of the calculation of the likelihood of a tree. (a) Data in the form of sequence alignment of length n . (b) One of three possible trees for the four taxa whose sequences are shown in (a). (c) The likelihood of a particular site, in this case site 5, equals the sums of the 16 probabilities of every possible reconstruction of ancestral states at nodes 5 and 6 in (b). (d) The likelihood of the tree in (b) is the product of the individual likelihoods for all n sites. (e) The likelihood is usually evaluated by summing the logarithms of the likelihoods at each site, and reported as the log likelihood of the tree. Modified from Swofford et al. (1996).

(a)	1	2	3	4	5	6	7	8	9	...	n
OTU1	A	A	G	A	C	T	T	C	A	...	N
OTU2	A	G	C	C	C	T	T	C	T	...	N
OTU3	A	G	A	T	A	T	C	C	A	...	N
OTU4	A	G	A	G	G	T	C	C	T	...	N



(c)

$$\begin{aligned}
 L_{(5)} = & \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ \text{A} - \text{A} \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ \text{A} - C \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ \text{A} - T \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ \text{A} - G \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) \\
 & + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ C - \text{A} \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ C - C \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ C - T \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ C - G \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) \\
 & + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ T - \text{A} \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ T - C \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ T - T \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ T - G \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) \\
 & + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ G - \text{A} \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ G - C \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ G - T \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right) + \text{Prob} \left(\begin{array}{c} C \\ \diagup \quad \diagdown \\ G - G \\ \diagdown \quad \diagup \\ C \end{array} \begin{array}{c} A \\ \diagup \quad \diagdown \\ \text{A} \\ \diagdown \quad \diagup \\ G \end{array} \right)
 \end{aligned}$$

(d) $L = L_{(1)} \times L_{(2)} \times L_{(3)} \times \dots \times L_{(n)} = \prod_{i=1}^n L_{(i)}$

(e) $\ln L = \ln L_{(1)} + \ln L_{(2)} + \ln L_{(3)} + \dots + \ln L_{(n)} = \sum_{i=1}^n \ln L_{(i)}$

Computing Maximum Likelihood Tree